2011

Genome engineering using DNA-binding proteins: zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs)

Deepak Reyon
Iowa State University

Genome engineering using DNA-binding proteins: zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs)

by

Deepak Reyon

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Drena Dobbs, Co-Major Professor
Edward Yu, Co-Major Professor
Vasant Honavar
Amy Andreotti
Gaya Amarasinghe

Iowa State University
Ames, Iowa
2011

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Over the past two decades, research groups in both academia and private industry have developed key technologies, including viral delivery vectors and engineered transposon-based or zinc finger protein-based nucleases, towards achieving the long-sought goal of therapeutic genome editing in humans. To date, Zinc Finger Nucleases (ZFNs) have been the most promising reagents for potential therapeutic applications in humans, but the recently characterized Transcription Activator Like Effector (TALE) proteins may soon change this status quo. Although it remains to be seen whether nucleases based on these proteins (TALENs) will be as broadly applicable and effective as ZFNs, based on initial reports, TALENs look very promising. Currently, the primary advantage of TALENs is that the DNA binding code for TALENs appears to be simple and robust, making their synthesis relatively simple.

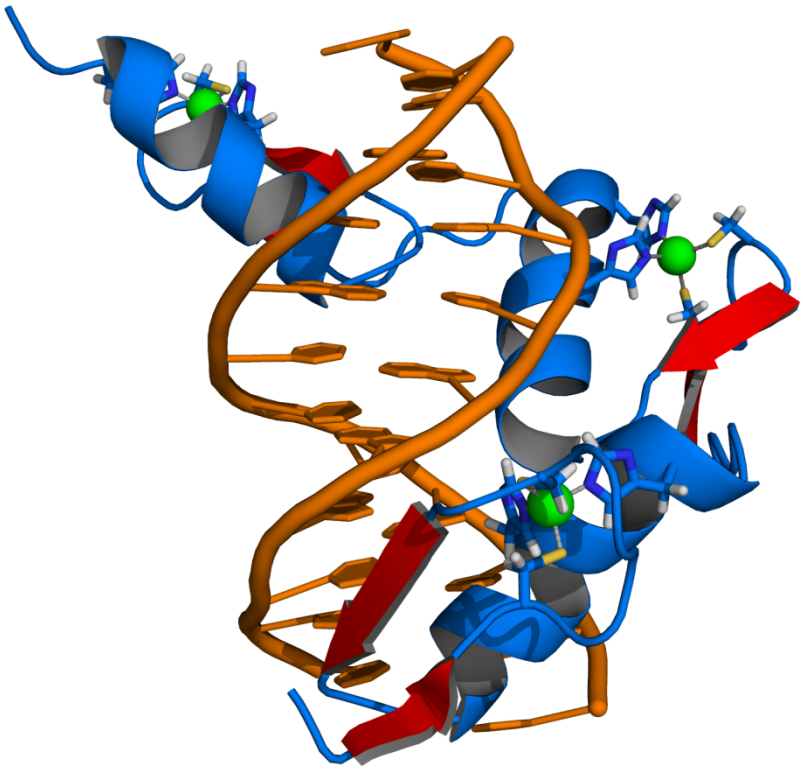
In this dissertation, I summarize advances made in the field of genome editing over the past decade and compare and contrast the currently available tools, focusing on ZFNs and TALENs. Specifically, I describe our efforts to make ZFN technology more accessible by designing and implementing models to help researchers choose target sites that are most amenable to targeting using ZFNs. Also, to help explore the potential of TALENs as tools for genome editing, I describe the development of a simple protocol to aid in constructing TALENs. As ZFNs become easier to use, and TALENs become more robust, the use of genome editing techniques as therapeutics appears poised to become reality in the near future.

CHAPTER 1. GENERAL INTRODUCTION

This dissertation describes a combined experimental and computational effort to genetically engineer and functionally characterize Zinc Finger Nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs), with the goal of developing improved tools for genomic modification.

BACKGROUND

In the years since Paul Berg first described a method (ligation) to join two pieces of DNA in 1972, scientists have been incrementally developing tools to modify genomes efficiently and precisely [1]. Although significant progress has been made in the ensuing four decades, the goal of using genome editing techniques as therapeutic tools in humans remains elusive. Because most of the progress has improved the “efficiency”, but not the “precision” of recombination, the problem that remains unsolved. The first experiments in genome engineering were conducted using homologous recombination (HR), which is a highly precise but extremely inefficient mechanism[2]. The next stage of development in genome engineering technology involved methods that were significantly more efficient (such as retroviruses, and transposons), but, also, less precise.



Credit: Thomas Splettstoesser

Figure 1.1. Cartoon representation of Zif268 (blue and red) in complex with DNA (orange), a zinc finger protein comprised of 3 modules. The zinc ion is represented as green spheres.

The field of genome engineering switched into high gear in 2001, when Bibikova and Carroll used zinc finger nucleases (ZFNs) [3, 4], first described by Chandrashekaran in 1996 [5], to induce a double-stranded breaks (DSB) into DNA in *Drosophila*, and observed a 1000-fold increase levels of HR. The concept that introduction of DSBs increases the efficiency of HR was first demonstrated by Maria Jasin in 1996, using I-SceI endonuclease [6]. Unlike the I-SceI endonuclease, a ZFN is composed of a "designable" ZFP fused to a non-sequence specific FokI endonuclease. The DNA binding specificity is provided by the ZFP, while the FokI nuclease provides the functionality. In theory, if one could design a ZFP that displays sufficient DNA binding specificity for a desired target site, in the context of an entire genome, genome editing would no longer be a pie in the sky. Unfortunately, designing and building highly specific ZFPs was not as

trivial as originally thought. Based on the crystal structure published by Pavletеich and Pabo in 1991 [7] (and improved by Fairall and Finch in 1993[8]), and a series of biochemical experiments performed by Berg, Miller, Klug, Roeder, Brown, Neuhaus, Wright and their colleagues [9-13], the hope was that the specificity of a ZFP was determined by the 4 residues that directly contact the bound DNA molecule. If the code for ZFP-DNA interaction were that simple, one could string together several ZFPs (each of which would specify 3 bps of DNA) and synthesize a zinc finger array (ZFA) with the desired sequence specificity. See Figure 1.1 and 1.2

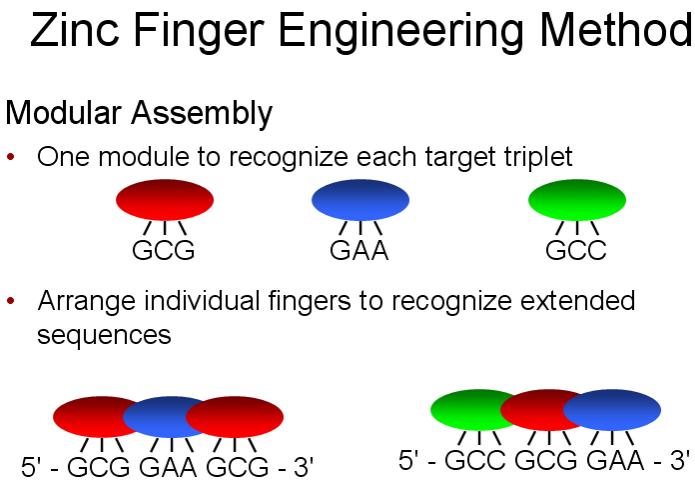
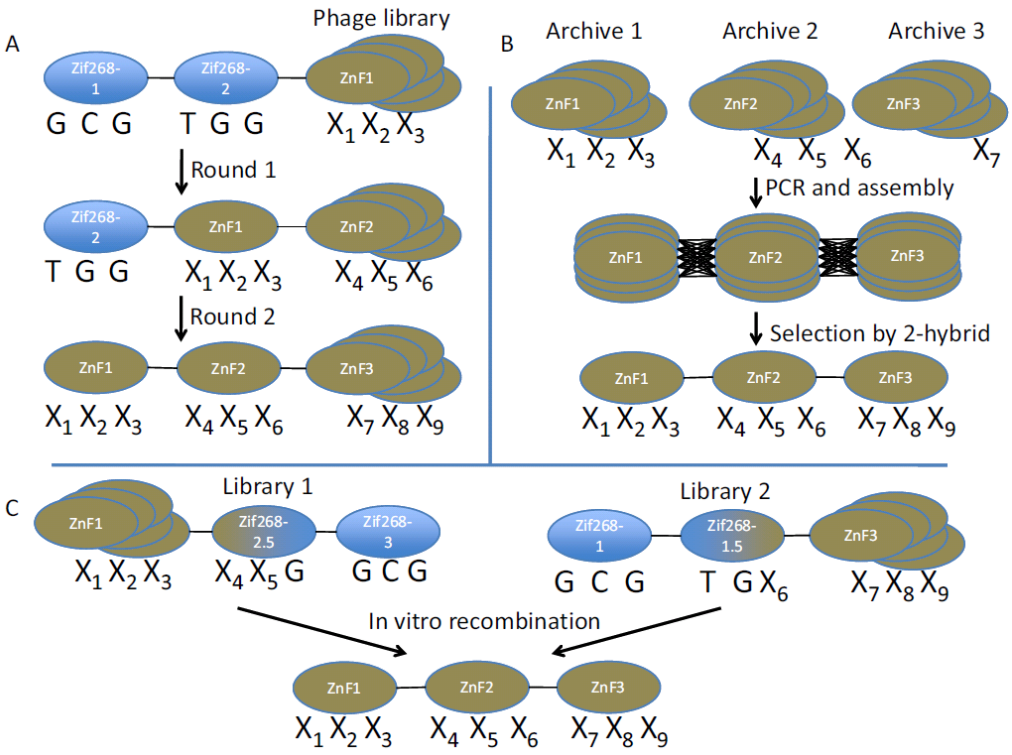


Figure 1.2. Schematic for modular assembly

This approach, commonly called “modular assembly” dominated the first wave of ZFN engineering; it was elegant and simple, but the failure rate of these enzymes was unacceptably high [14]. The most plausible explanation for the high failure rates was that the “recognition code” describing the ZFN–DNA interaction wasn’t as simple as initially thought. Simply stringing together ZFPs that have been designed to bind a given 3-bp target site isn’t sufficient as the behavior of a single ZFP is different from a ZFP that is part of a ZFA. The conundrum is that when the 1-to-1 ZF-DNA binding code breaks down, the design problem becomes exponentially more difficult. The set reagents required to build ZFPs to a 9-bp DNA target site increases from 64 to 262144 as we would have to design a protein to every 9-bp target site rather than every 3-bp target site.

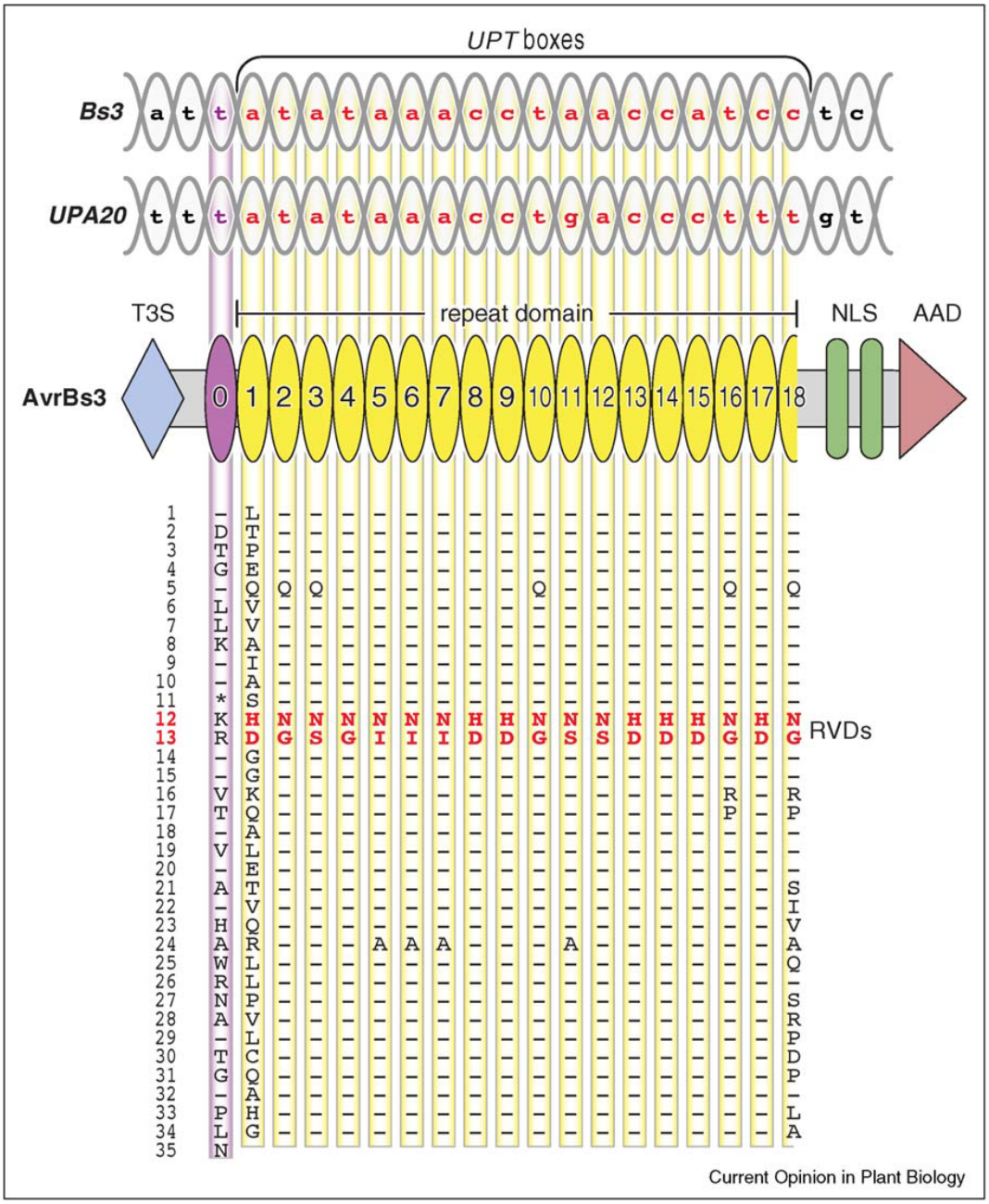
The current state-of-the-art methods for ZFN engineering take into account that the DNA sequence recognized by a ZFP depends on the entire protein sequence, rather than a code simply defined by individual ZF modules that are pasted together. In other words, the current methods take into account the fact that the *context* of an individual ZF module (its neighboring modules, as well as its position in a linear array of several modules), can dramatically impact its DNA recognition properties. These newer methods, e.g., Sangamo, OPEN, Greismann, and Isalan (reviewed by Davis and Stokoe [15]) involve making large libraries of potential ZFNs, followed by several rounds of selection to identify candidate ZFNs with high relative affinity for the desired target DNA sequence. These methods have improved the success rates of ZFNs, but are much more technically demanding than their predecessor, modular assembly [16].



Credit: Davis and Stokoe [15]

Figure 1.3. Summary of selection methods used to design ZFPs. A. Greismann method. **B.** Oligomerized pool Engineering (OPEN). **C.** Isalan, Klug method.

In 2009, two groups, Boch and Bogdanove, reported that a surprisingly simple code governs the DNA recognition properties of a family of bacterial DNA binding proteins called Transcription Activator-Like Effectors (TALEs) [17, 18]. This report kicked the field into overdrive. Although it is still too early to rigorously evaluate the utility of TALEs (and judge their advantages vs. ZFNs), the frenzy created by this discovery is undeniable, and based on results from initial published reports, they do not suffer from several drawbacks of ZFPs. Most importantly, TALENs do seem to be truly modular, with a 1–to-1 correspondence between a repeat variable di-residue (RVD) in the TALE protein sequence and a single nucleotide in the DNA sequence (in which NI specifies A, NN specifies G, NG specifies T, and HD specifies C) [19]



Credit: Bogdanove et al. [19]

Figure 1.4. Structure of a TALE.

Benefiting from the lessons learned by the zinc finger community, TALE engineering has progressed in leaps and bounds over the past 2 years (~ 20 high-profile papers describing examples of TALEN-mediated genome editing since the seminal 2009

papers were published). Furthermore, the construction of TALEs is significantly easier than ZFPs. In the next section I've summarized the currently available methods.

METHODS FOR ASSEMBLING TALENs.

1. Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes [20].

   This assembly method described by Li et al., creates 8 sets of RVD containing domains (total of 32) via PCR. Each of these 8 sets are designed to have different overhangs when digested using type IIS restriction enzyme BsmBI, such that, they would ligate in a fixed order - 1 through 8. A potential drawback of this assembly method is that the lengths of the constructs have to be multiples of 8. However, an easy work around is to create a compatible end on the final unit via PCR. The TAL constructs are then cloned into a repeat-deficient pAvrXa7-FN scaffold which has 288-aa on the N-Term and 196-aa on the C-term.

2. Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting [21]

   This method described by Cermak et al. uses the golden gate strategy, which involves a series of digestions using type IIs restriction enzymes to generate unique overhangs such that a single ligation step would yield the final construct. The advantage of the method is that it is fast. A TAL construct can be generated as little as 5 days. The disadvantage to this method is that it is complicated and there is a significant startup cost associated – about 60 different constructs. The TALE constructs are then cloned into vectors previously described in [22] that provide the N term (287 AA) and C term (230 AA) of the TAL.

3. Assembly of custom TALE-type DNA binding domains by modular cloning.[23]

   This method described my Morbitzer et al., also utilizes the golden gate technology. The main difference, as described in the manuscript, is that this method can be used

to make constructs of fixed length - 10, 17 and 20. This method also uses only 1 type IIS restriction enzyme – BsaI. Similar to Cermak et al., users need to acquire at least 52 constructs before they proceed to build TALEs. The TAL constructs are then cloned into an *avrBs3* scaffold.

4. Targeted gene disruption in somatic zebrafish cells using engineered TALENs. [24]
This method described by Sander et al. is perhaps the simplest of the lot, and also the most time consuming. This essentially involves a series of digestions and ligations (see figure 1.5).

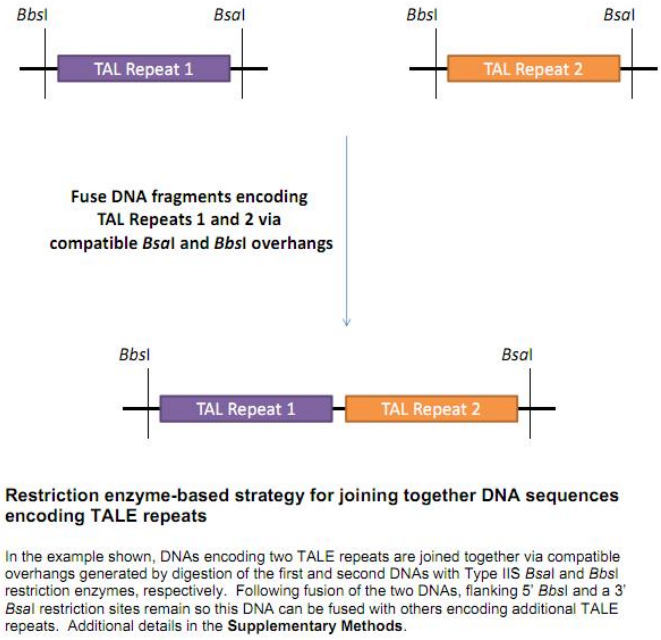
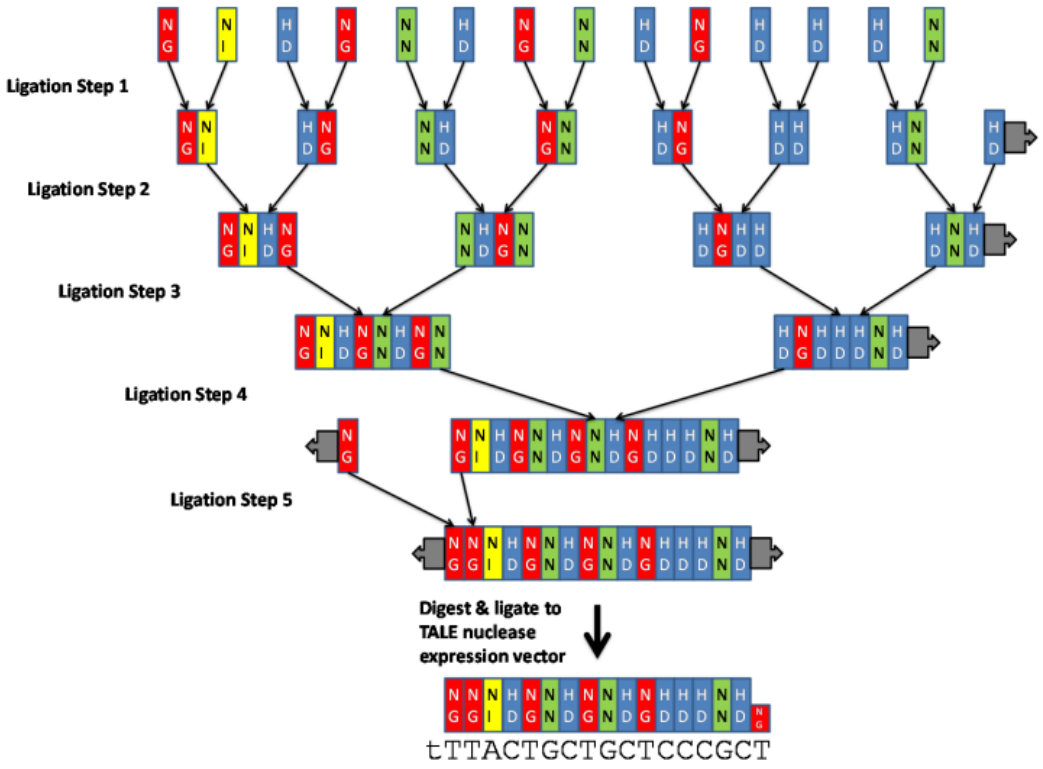


Figure 1.5. Schematic of the TAL units used in the assembly described in Sander et al. (Figure from [24]).

The assembly method is based on the framework published by Miller et al. To summarize: Miller et al.[25] showed that TAL scaffolds that include 135AA (153-288 from the endogenous TAL13 TALE) on the N-term and 63 AA (715 – 777 from the endogenous TAL13 TALE) on the C-term resulted in the most effective nucleases. Although, the assembly method isn't as fast as the rest of the methods

summarized here, it has the advantage of using the framework that has been used most successfully. Evidence of genome editing using TALEN designed using this framework have been shown to work in a variety of organisms including zebrafish, *C. elegans*, rat, human somatic cells, and human pluripotent stem cells. The assembly strategy is summarized in the following figure.



Schematic overview of assembly strategy used to engineer TALE repeat array for TALE nuclease #1257

A series of serial ligation steps were used to assemble together the 16 TALE repeats of TALE nuclease #1257 using the restriction enzyme-based strategy detailed in **Supplementary Figure 8**. TALE repeats are shown as colored rectangles with the RVDs abbreviated as two letters. Modified sequences on the 5' and 3' ends of the DNA encoding the N- and C-terminal TALE repeats, respectively, that are required for cloning into the final TALE nuclease expression vector are illustrated as grey colored arrow boxes. Cloning into the TALE nuclease expression vector in the last ligation step creates a plasmid that expresses the TALE repeat array fused to the C-terminal 0.5 TALE repeat (shown as a smaller colored rectangle). See **Supplementary Methods** for additional details.

Figure 1.6. TAL assembly process described in Sander et al. (figure from [24]).

Another confounding factor in TALEN engineering is the lack of constraints while choosing target sites. Sander et al. describe an addendum to the ZiFiT software

from the Dobbs, Joung and Voytas labs to aid in the identification of target sites. http://zifit.partners.org/ZiFiTBeta/

5. Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription. [25]

   This assembly method described by Zhang et al., differs for the rest of the methods in that the unique overhangs are introduced via PCR. Rather than starting off with a set of modules that can be cyclically digested and ligated, in this method 12 separate PCR reactions are performed on each of the four RVD modules. These PCR products are then digested and ligated to form 4mers. These 4mers are then PCR amplified and digested. These fragments are then ligated to assemble a TAL array of length 12. This method should be expandable to generate longer TAL, but, as published, 12mer are the only option. Another potential problem is the errors that could be introduced during PCR. The 12mers built using this method is cloned into a TAL scaffold that has the first 48-aa deleted from the N-term and the C-term is also truncated to a length of 68-aa.

6. Assembly of designer TAL effectors by Golden Gate Cloning. [26]

   This assembly method, also based on the golden gate technology, is very similar to the method described by Feng Zhang et al., sans the PCR amplification. This method is potentially superior to Zhang et al, because it can be used to synthesize TALEs that bind 17mers rather than 12mers. A potential drawback is that this method can be used to build only 17mers. Furthermore, the upfront cost is considerable as there are 17 * 4 modules + backbones. The 17mer constructs are then cloned into an *avrBs3* scaffold.

7. Transcriptional Activators of Human Genes with Programmable DNA-Specificity. [27]

   This method described by Geissler et al., is also based on the golden gate strategy and involves the construction of 6mer (or less) subunits that are than ligated together

to form a full length TALEs. As this method assembles subunits as 6mers so, only 6*4 units, 5 termination units, and 6 assembly vectors are required. Though this assembly method is not as fast as Cermak et al., it has the advantage of being simple and easy to practice. The constructs assembled using this method is then cloned into a Hax3 scaffold.

Of these methods, Cermak et al., [21] is probably the fastest among the flexible (any length TALs) frameworks. And, the Miller et al.[25] framework has the most number of published gene editing successes. Ultimately, it will not matter which assembly method is better because the field will benefit either way. The sooner we solve the problem of acquiring high quality reagents, the sooner we can start dealing with other inherent problems that accompany any biological system.

OVERALL GOAL AND RESEARCH AIMS

The overall goal of this dissertation is to develop improved tools for genomic modification. Specifically, my research is focused on three complementary aims: 1) to develop and implement computational tools and resources to facilitate the design and evaluation reagents made using the two dominant technologies for genome engineering at present, Zinc Finger Nucleases (ZFNs) and Transcription Activator-Like Effector Nucleases (TALENs); 2) to apply these tools in the design of novel ZFN and TALEN reagents to target specific DNA sequences; and 3) to experimentally evaluate the functional activities of the designed ZFNs and TALENs, both *in vitro* and *in vivo,* in several model genetic organisms. An important aspect of this study, which has only very recently become possible, is a comparison of the relative advantages and disadvantages of ZFN *vs.* TALEN technologies for genome editing applications.

ZFNs have been studied for over 20 years, and, hence, have an inherent advantage because we have a much deeper understanding of their three-dimensional structures, mode of DNA binding, and requirements for activity. TALENs, on the other hand, are so new that we do not yet have a complete three-dimensional structure and can only speculate about their mode of DNA binding. TALENs seem to function at least as well as

ZFNs, however, and offer one significant advantage over ZFNs: they are very easy to design and synthesize. The major drawback of the TALEN technology is that relatively few TALEN enzymes have been functionally tested, and the activities of novel engineered TALENs may be unpredictable. It remains to be seen which one of these technologies will dominate the field of genome engineering in the future.

DISSERTATION ORGANIZATION

The dissertation has five chapters and four appendices. ***Chapter 1*** is a general introduction that focuses on comparison of the various methodologies for engineering ZFNs and TALENs as tools for genomic modification, ***Chapter 2*** is a paper published in BMC *Bioinformatics* in 2010, on which I am co-first author. This manuscript describes the development of a Naïve Bayes classifier to help researchers identify and rank potential ZFN target sites that are most amenable to targeting using ZFNs generated using the most successful publically available protocol – OPEN from the Joung Lab. ***Chapter 3*** is a paper published in *BMC Genomics* in 2011, on which I am first and corresponding author. This paper presents all potential target sites within the context of the entire genomes of several model organisms. To further assist researchers with picking target sites we also score each target sites using the Naïve Bayes classifier described in ***Chapter 2*** and a "uniqueness" score. ***Chapter 4*** is a paper published in the *Nature Biotechnology* in 2011, describing the construction of TALE nucleases using the framework described in Miller et al [25]. My contribution to this work included construction of the TALE nucleases that were tested in Zebrafish. ***Chapter 5*** is a co-first-authored paper currently under review for publication in *Nature Protocols*, describing, in detail, the protocol used to assemble TALENs used in Sander et al. [24]. In this paper we also describe the modification of the ZiFiT software [28] to include the capability to identify TALEN target sites. ***Chapter 6*** includes general conclusions, a summary of the major contributions of this dissertation research, and suggested future directions. ***Appendix A*** is a paper published in *Nature Methods* in 2011, in which we describe a new method to construct designer zinc finger proteins called context dependent assembly (CoDA). Prior to the description of this method the two prevalent methods were modular assembly,

which does not take context into account and as a result was easy to practice, but also had an unacceptably high failure rate; and selection based methods that had a higher success rate, but were significantly more difficult to practice. CoDA is a hybrid. In a nutshell, CoDA involves the assembly of 2 finger modules that were identified in three finger selections. My contributions to this paper include the identification of all CoDA target sites in several model organisms and the calculation of varies statistics to evaluate the effectiveness of CoDA. After the publication of the paper I performed some selections to improve the targeting range of CoDA. This data is summarized in Table 1.1.

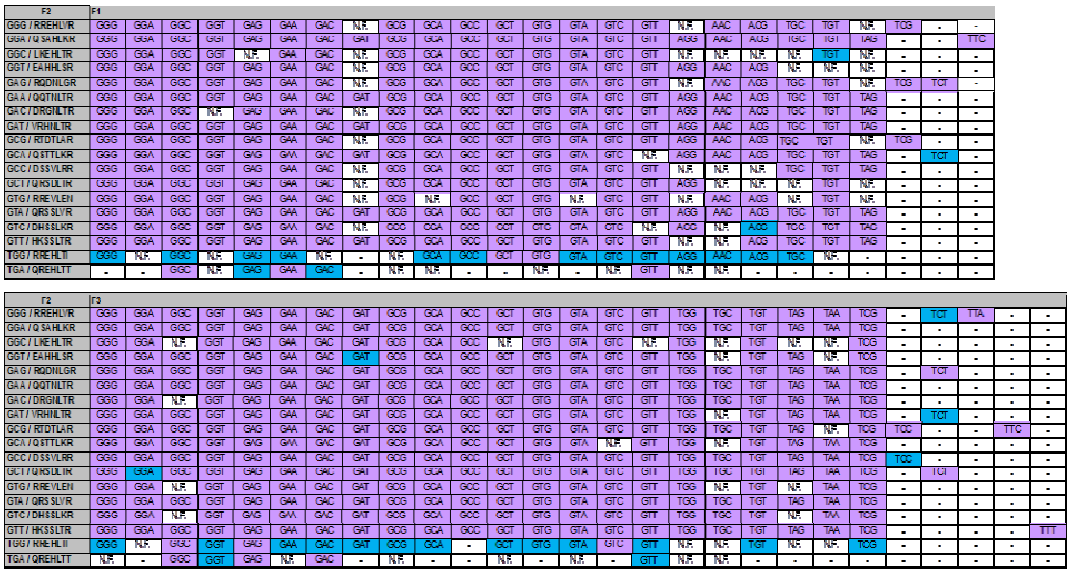


Table 1.1. Published (Purple) and new unpublished (Blue) CoDA Zinc Finger Units

Appendix B is a paper published in *Nucleic Acids Research* in 2010 (web server issue) presenting ZiFiT 2.0. ZiFiT is a web server that identifies target sites for zinc finger proteins designed using the publically available methods – modular assembly, OPEN, and CoDA. ZiFiT is dynamically linked to ZiFDB, which is a repository for previously designed ZFPs and NCBI BLAST to aid in identification of unique sites. Recently, ZiFiT was updated to include scores generated using ZiFOpT (the classifier described in chapter 2) and also identify TALEN target sites (described in Chapter 5). ***Appendix C*** includes Supplemental Data for papers in Chapters 4, 5, and appendix A. ***Appendix D*** is my *Curriculum vitae.*

REFERENCES

1. Jackson, D.A., R.H. Symons, and P. Berg, *Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of Escherichia coli.* Proc Natl Acad Sci U S A, 1972. **69**(10): p. 2904-9.
2. Folger, K.R., et al., *Patterns of integration of DNA microinjected into cultured mammalian cells: evidence for homologous recombination between injected plasmid DNA molecules.* Mol Cell Biol, 1982. **2**(11): p. 1372-87.
3. Bibikova, M., et al., *Stimulation of homologous recombination through targeted cleavage by chimeric nucleases.* Mol Cell Biol, 2001. **21**(1): p. 289-97.
4. Bibikova, M., et al., *Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases.* Genetics, 2002. **161**(3): p. 1169-75.
5. Kim, Y.G., J. Cha, and S. Chandrasegaran, *Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain.* Proc Natl Acad Sci U S A, 1996. **93**(3): p. 1156-60.
6. Jasin, M., *Genetic manipulation of genomes with rare-cutting endonucleases.* Trends Genet, 1996. **12**(6): p. 224-8.
7. Pavletich, N.P. and C.O. Pabo, *Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A.* Science, 1991. **252**(5007): p. 809-17.
8. Fairall, L., et al., *The crystal structure of a two zinc-finger peptide reveals an extension to the rules for zinc-finger/DNA recognition.* Nature, 1993. **366**(6454): p. 483-7.
9. Berg, J.M., *Proposed structure for the zinc-binding domains from transcription factor IIIA and related proteins.* Proc Natl Acad Sci U S A, 1988. **85**(1): p. 99-102.
10. Brown, D.D., *The role of stable complexes that repress and activate eukaryotic genes.* Philos Trans R Soc Lond B Biol Sci, 1984. **307**(1132): p. 297-9.
11. Miller, J., A.D. McLachlan, and A. Klug, *Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes.* Embo J, 1985. **4**(6): p. 1609-14.
12. Lee, M.S., et al., *Three-dimensional solution structure of a single zinc finger DNA-binding domain.* Science, 1989. **245**(4918): p. 635-7.
13. Neuhaus, D., et al., *Solution structures of two zinc-finger domains from SWI5 obtained using two-dimensional 1H nuclear magnetic resonance spectroscopy. A zinc-finger structure with a third strand of beta-sheet.* J Mol Biol, 1992. **228**(2): p. 637-51.
14. Ramirez, C.L., et al., *Unexpected failure rates for modular assembly of engineered zinc fingers.* Nat Methods, 2008. **5**(5): p. 374-5.
15. Davis, D. and D. Stokoe, *Zinc finger nucleases as tools to understand and treat human diseases.* BMC Med, 2010. **8**: p. 42.
16. Sander, J.D., et al., *Predicting success of oligomerized pool engineering (OPEN) for zinc finger target site sequences.* BMC Bioinformatics, 2010. **11**: p. 543.

17. Boch, J., et al., *Breaking the code of DNA binding specificity of TAL-type III effectors.* Science, 2009. **326**(5959): p. 1509-12.
18. Moscou, M.J. and A.J. Bogdanove, *A simple cipher governs DNA recognition by TAL effectors.* Science, 2009. **326**(5959): p. 1501.
19. Bogdanove, A.J., S. Schornack, and T. Lahaye, *TAL effectors: finding plant genes for disease and defense.* Curr Opin Plant Biol, 2010. **13**(4): p. 394-401.
20. Li, T., et al., *Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes.* Nucleic Acids Res, 2011. **39**(14): p. 6315-25.
21. Cermak, T., et al., *Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting.* Nucleic Acids Res, 2011. **39**(12): p. e82.
22. Christian, M., et al., *Targeting DNA double-strand breaks with TAL effector nucleases.* Genetics, 2010. **186**(2): p. 757-61.
23. Morbitzer, R., et al., *Assembly of custom TALE-type DNA binding domains by modular cloning.* Nucleic Acids Res, 2011. **39**(13): p. 5790-9.
24. Sander, J.D., et al., *Targeted gene disruption in somatic zebrafish cells using engineered TALENs.* Nat Biotechnol, 2011. **29**(8): p. 697-8.
25. Miller, J.C., et al., *A TALE nuclease architecture for efficient genome editing.* Nat Biotechnol, 2011. **29**(2): p. 143-8.
26. Weber, E., et al., *Assembly of designer TAL effectors by Golden Gate cloning.* PLoS One, 2011. **6**(5): p. e19722.
27. Geissler, R., et al., *Transcriptional activators of human genes with programmable DNA-specificity.* PLoS One, 2011. **6**(5): p. e19509.
28. Sander, J.D., et al., *Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W599-605.

CHAPTER 2. PREDICTING SUCCESS OF OLIGOMERIZED POOL ENGINEERING (OPEN) FOR ZINC FINGER TARGET SITE SEQUENCES

Jeffry D. Sander*, Deepak Reyon*, Morgan L. Maeder, Jonathan E. Foley, Stacey Thibodeau-Beganny, Xiaohong Li, Fengli Fu, Daniel F. Voytas, J. Keith Joung & Drena Dobbs. *BMC Bioinformatics*. 2010 Nov 2.

*Co-first authors

ABSTRACT

Background:

Precise and efficient methods for gene targeting are critical for detailed functional analysis of genomes and regulatory networks and for potentially improving the efficacy and safety of gene therapies. Oligomerized Pool ENgineering (OPEN) is a recently developed method for engineering C2H2 zinc finger proteins (ZFPs) designed to bind specific DNA sequences with high affinity and specificity *in vivo*. Because generation of ZFPs using OPEN requires considerable effort, a computational method for identifying the sites in any given gene that are most likely to be successfully targeted by this method is desirable.

Results:

Analysis of the base composition of experimentally validated ZFP target sites identified important constraints on the DNA sequence space that can be effectively targeted using OPEN. Using alternate encodings to represent ZFP target sites, we implemented Naïve Bayes and Support Vector Machine classifiers capable of distinguishing "active" targets, i.e., ZFP binding sites that can be targeted with a high rate of success, from those that are "inactive" or poor targets for ZFPs generated using current OPEN technologies. When evaluated using leave-one-out cross-validation on a dataset of 135 experimentally validated ZFP target sites, the best Naïve Bayes classifier, designated ZiFOpT achieved 87% overall accuracy and specificity^{+} of 90%, with an AUC of 0.89.

The performance of ZiFOpT on a completely independent (blind) test set of 66 validated ZFP target sites was comparable (88% accuracy; 92% specificity^{+}).

Conclusion:

ZiFOpT, a machine learning classifier trained to identify DNA sequences amenable for targeting by OPEN-generated zinc finger arrays, can guide users to target sites that are most likely to function successfully *in vivo*, substantially reducing the experimental effort required.

BACKGROUND

Zinc finger (ZF) DNA binding proteins can be used to target functional protein domains to specific regions in complex genomes. For example, zinc finger nucleases (ZFNs) have tremendous potential for introducing site-specific gene knockouts or gene targeting events with high efficiency in various cell types including human [1, 2]. A ZFN consists of two zinc finger proteins (ZFPs) each fused to a monomeric *Fok*I nuclease domain. When the ZFPs co-locate to adjacent sequences within the genome, the nuclease monomers are able to dimerize, generating an active nuclease that cleaves the double-stranded DNA at the target site. In the presence of exogenous donor DNA, genetic material may be exchanged through repair by homologous recombination; alternatively, the break may be repaired by non-homologous end joining, which is an error-prone mechanism that commonly results in knockout mutations [3, 4]. To date, ZFNs have been used to manipulate endogenous genes in several organisms, e.g., tobacco, maize, fruit fly, zebrafish, rats, and human [5-14], and are being evaluated in human clinical trials, including gene therapies to treat AIDS [15-17].

Zinc finger DNA binding domains, especially the C2H2 class of zinc fingers, have been exploited for performing targeted genome modification because they can be engineered to bind a wide range of desired DNA sequences. Each individual C2H2 zinc finger consists of an α-helix (the DNA "recognition helix") and a β-hairpin, stabilized by a single zinc ion coordinated through interactions with cysteine and histidine residues. Individual ZFs recognize and bind specific triplet DNA sequences through base-specific contacts within the major groove of double-stranded DNA [18]. Extended DNA

sequences can be targeted by joining together several ZF domains [1, 19]. In the traditional "modular assembly" approach to ZF engineering, one ZF domain is identified for targeting each nucleotide triplet sequence [20-22]. Three-finger arrays generated using modular assembly have been shown to bind specifically to their 9-bp targets *in vitro* [23]. However, several studies have demonstrated that many ZF arrays fail to function successfully *in vivo* [7, 24-26]. These failures have been attributed to insufficient affinity of engineered arrays [25], variations in target specificity or affinity due to inter-finger context dependencies [24, 27].

Compared with modular assembly, ZFP design using the recently developed Oligomerized Pool ENgineering (OPEN) method has been reported to provide higher *in vivo* success rates, particularly for zinc finger nucleases (ZFNs) [7, 8]. For constructing ZFPs that recognize 9-bp targets, the OPEN method involves combinatorial assembly and subsequent selection of fingers from three pre-constructed pools, each of which contains up to 95 different engineered ZF recognition helix "solutions" for a chosen DNA triplet [7, 28]. Currently, pools are available for all 16 GNN triplets and several of the TNN triplets for each position in a three-finger array [7]. ZFNs generated by OPEN have been used to target genes in tobacco, zebrafish, and human cells with high efficiency [7-9].

Because using the OPEN procedure requires investment of time and effort and because there are often numerous potential targetable sites in any given gene, it is desirable to focus experiments on target sites that are most likely to yield functional ZFPs. For example, there are 315,186 OPEN ZFN sites in the protein encoding regions of the zebrafish genome (an average of 10.8 sites per transcript). While OPEN often generates ZFPs that function well in a bacterial two-hybrid (B2H) reporter system [7, 8], it does not have a 100% success rate. Thus, to reduce the experimental effort involved in applying the OPEN procedure, we sought to develop a computational approach to identify the "best" targets, i.e., those most likely to be successfully targeted by OPEN, from among the relatively large number of theoretically "targetable" ZFP sites that may exist for any chosen gene or genomic region of interest.

In this study, we demonstrate that sequence characteristics of ZFP target sites, when used as input to Naïve Bayes or Support Vector Machine (SVM) classifiers, can be

used to reliably predict whether a specific DNA sequence will (or will not) be successfully targeted by OPEN. The performance of these classifiers on two experimentally validated datasets of ZF target sites suggests that their use could substantially reduce the experimental effort required to generate a functional ZFN using the OPEN method.

RESULTS

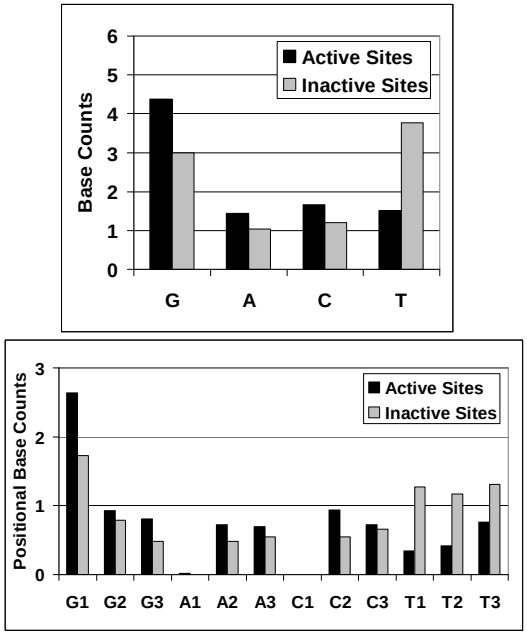
Results from several groups [24, 29, 30] have suggested that ZFP recognition sites with a high purine nucleotide content, especially those containing several GNN-triplets, more frequently correspond to "active" targets for zinc finger proteins generated using modular assembly. To investigate whether such potential biases could be exploited to identify optimal sequences for ZFP targeting using OPEN, we analyzed sequence and base composition characteristics of sites targeted by this method.

For this study, we first generated an experimentally-validated dataset, ZFTS135, consisting of 135 9-bp target sites for which OPEN did or did not successfully yield ZFPs. ZFTS135 includes 53 ZF target sites from recently published OPEN experiments [7, 8] and 82 OPEN ZF target sites which we report here for the first time. Each target site in the dataset was assigned a class label of either "active" (79%) or "inactive" (21%). “Active” target sites were those yielding at least one ZFP that showed DNA-binding activity in a well-validated bacterial two-hybrid (B2H) reporter assay (defined as the ability to activate transcription by three-fold or more, a level previously shown to identify ZF arrays that possess high affinity and high specificity for their cognate DNA binding site [7, 28]). “Inactive” target sites were those that failed to yield a ZFP that showed activity in the B2H reporter assay. All 135 functionally validated ZFP target sites and their assigned labels are provided in Supplemental Table 1.

Figure 2.1 presents analyses of the sequence and base composition characteristics of ZFP target sites in the ZFTS135 dataset. The average number of times each base occurs in active and inactive targets is shown in Figure 2.1 A. On average, active sites contain more guanines and fewer thymines than inactive targets. Because OPEN ZF finger pools are available exclusively for GNN and TNN triplet sub sites at present, total

guanine and thymine counts are inflated, compared to adenine and cytosine counts. To account for this, as well as the fact that specific bases, when located in different positions within a triplet sub site, may preferentially contact different amino acids, the average base occurrences were calculated for each position within the triplets (Figure 2.1 B). This analysis identified thymine frequency, at any position within a triplet, as the primary difference between active and inactive target sites. Guanine, adenine, and cytosine typically appear more frequently in active sites than in inactive sites, compensating for the decrease in thymine content.

a) **b)**



c)

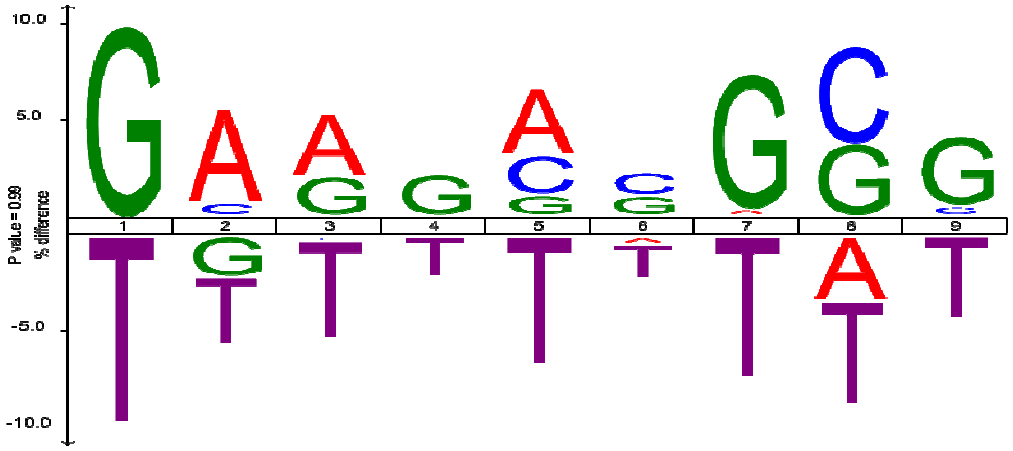


Figure 2.1 | Base composition differs in active versus inactive ZFP target sites.
A) Total *base counts* for active and inactive ZFP target sites (from ZFTS135, a dataset of 135 experimentally validated 9-bp target sites, see Supplemental Table 1) reveal that variation in the average frequency of each base differentiates active versus inactive target sites. The total number of G and T residues relative to A and C is inflated because currently available OPEN pools are designed to target GNN and TNN triplets. **B)** *Positional base counts*, i.e., average base counts for each position within target site triplets (1st, 2nd, 3rd), suggest that thymine bases negatively impact ZFP binding at all three positions. **C**) An iceLogo [31] generated from ZFTS135 illustrates the difference in percentage composition of nucleotides at each position, from 1 – 9 (5' to 3'), between the positive class and the entire dataset. For example, 78% of all sites in ZFTS135 have a G in position 1, whereas 88% of all *active* sites have a G at position 1, resulting in a difference of 10%. Positive difference values indicate that, on average, the indicated bases are favored at those positions in active sites; negative difference values indicate that the indicated bases are disfavored. These position-specific differences in percentage composition also support the conclusion that thymine bases tend to occur in inactive targets (i.e., they have large negative propensities).

Differences in base composition at each position within active 9-bp target sites were also analyzed. As shown in Figure 2.1 C, thymine is generally disfavored in active target sites, with strong negative propensities in the 1st and 7th positions of active target sites. Other residues showed marginally positive propensities in most positions. Because available OPEN reagents are currently limited to those that target GNN and TNN triplets [7] (and one ANN triplet; M. Maeder & J.K. Joung, unpublished data), it is not possible to evaluate the significance of the relatively low percentage of adenine and cytosine residues in positions 1, 4 and 7.

Taken together, the results of these analyses suggested that base composition biases in active versus inactive ZFP target sites could be exploited by machine learning classifiers to

predict whether a specific DNA sequence can be targeted successfully using the OPEN procedure. Machine learning classifiers that use a string of sequence identities as input have been successfully applied to a variety of problems, including protein functional site classification [32-35]. Because several different machine learning classifiers we tested gave comparable results (data not shown), here we present representative results obtained using two types of classifiers: Naïve Bayes and support vector machines (SVMs).

We compared classifiers trained using three different target site sequence encodings: i) *sequence identity*: 9 nucleotide identities corresponding directly to the target site sequence; ii) *base counts*: 4 numerical values representing the overall base counts of G,A,C,T in the target site; iii) *positional base counts*: 12 numerical values encoding the position-specific base composition of the target site (see Methods for details).

Table 2.1 | Performance of classifiers in predicting active OPEN target sites.

Classifier	Target site encoding	Correlation Coefficient	Accuracy %	Specificity^{+} %	Sensitivity^{+} %	ROC AUC
Naïve Bayes	Sequence Identity	0.61	87	90	94	0.89
	Base Counts	0.57	87	89	94	0.79
	Positional Base Counts	0.59	87	88	97	0.84
SVM	Sequence Identity	0.48	84	86	95	0.76
	Base Counts	0.54	85	89	92	0.78
	Positional Base Counts	0.63	88	90	95	0.84

Table 2.1 summarizes performance statistics for Naïve Bayes and SVM classifiers tested using the three different target site encodings and evaluated using leave-one-out cross-validation. In these experiments, classifiers were optimized for correlation coefficient, which is an indicator of how effectively a classifier identifies both positive (active) and negative (inactive) instances. All classifiers achieved correlation coefficients between 0.48 and 0.63, with accuracies $\geq$ 84%. For the practical application of identifying target sites for ZFPs that provide the greatest chance of success (for cases in which several potential target sites are available), it is appropriate to choose a classifier with a high specificity^{+} value, i.e., one that predicts a smaller number of "active" sites with higher confidence, rather than a high correlation coefficient *per se.*

The receiver operating characteristic (ROC) curves in Figure 2.2 illustrate the tradeoffs between true positive rate (TPR), i.e., the percentage of active target sites *correctly* predicted as such, and false positive rate (FPR), i.e., the percentage of inactive sites *incorrectly* predicted to be active, for the different target sequence encodings. Using the base counts and positional base counts encodings, the Naïve Bayes and SVM classifiers gave similar results. Based on the Area Under the Curve (AUC) of the ROC curves, the best overall results were obtained using the sequence identity encoding with the Naïve Bayes classifier (AUC = 0.89), which slightly outperformed the best SVM classifier (AUC = 0.84). We designate the sequence-based Naïve Bayes classifier, ZiFOpT, for Zinc Finger OPEN Targeter.

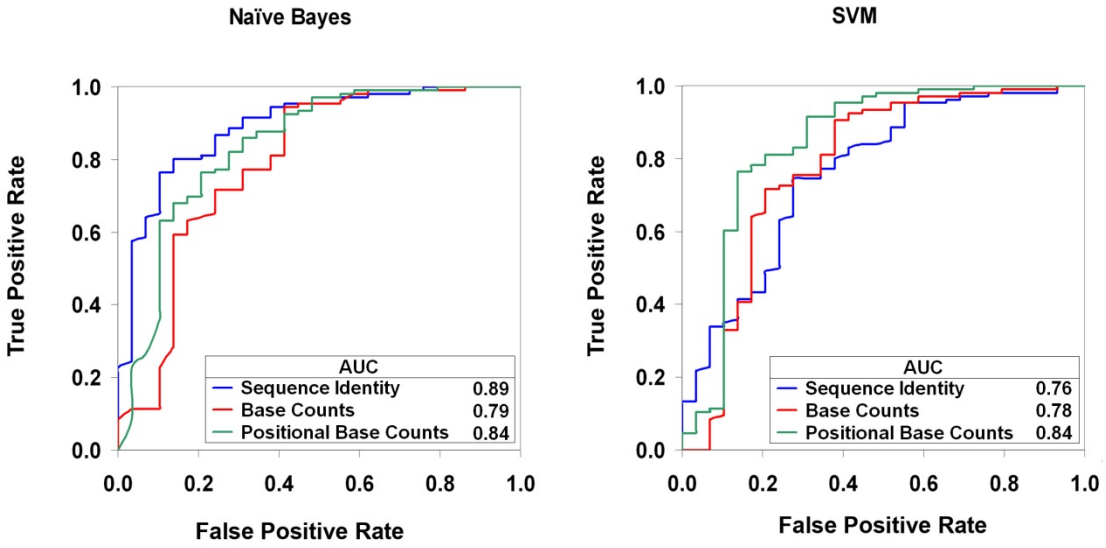


Figure 2.2 | Receiver Operating Characteristic (ROC) curves for Naïve Bayes and SVM classifiers.

To ensure that the performance of ZiFOpT on ZFTS135 was not over-estimated due to over-fitting, we generated a second completely independent data set of experimentally validated ZFP target sites. ZFTS66 consists of 66 9-bp target sites that were chosen by experts, unaided by our classifier, as ideal candidates for OPEN selection (see Supplemental Table 2). Active ZFPs were found for 58 of the 66 sites tested. Using ZiFOpT, we were able to accurately classify 88% of the sites, with 92% specificity^{+} and 95% sensitivity^{+} (Table 2, bottom row). Thus, ZiFOpT performs reliably when evaluated either by cross-validation or on an independent “blind” test set.

It is possible to rank the predicted active sites using a confidence score derived from the posterior probability returned by ZiFOpT (see Methods). As shown in Table 2, using a confidence score cutoff ≥ 6 results in improved accuracy, specificity^{+} and sensitivity^{+}.

Table 2.2 | Performance of ZiFOpT on an independent test set (ZFTS66)

Confidence Score	Accuracy %	Specificity^{+} %	Sensitivity^{+} %
≥ 6	94	94	100
< 6	67	75	81
Optimized for CC	88	92	95

Due to the large number of potential OPEN target sites for most genomic targets of interest, it is desirable to identify a subset of target sites with the greatest chance of success. Currently, OPEN pools are available for 26 triplets in position 1, 21 triplets in position 2, and 23 triplets in position 3 of a 3-finger ZFP. Hence OPEN can, in theory, target 12,558 distinct sites. Because 415 of these sites are not targetable due to *dam* or *dcm* methylation, 12,143 distinct 9-bp ZFP target sites are currently targetable. The ZiFOpT classifier, when optimized for correlation coefficient, predicts that 8,412 (69%) of these sites will be active target sites. For ZF nuclease sites, which consist of two ZF array sites, OPEN can theoretically target a total 147,452,449 distinct nuclease sites (assuming a fixed number of nucleotides between the arrays). ZiFOpT predicts that only 70,761,744 (48%) of these nuclease sites will have two active sites.

An analysis of recently published OPEN ZFN sites in zebrafish [8] illustrates the value of ZiFOpT in reducing the experimental effort required to target a large number of genomic transcripts. In the previous study, at least one potential OPEN nuclease site was identified within the first three coding exons in ~86% of zebrafish transcripts [8]. As shown in Table 3, using a classification threshold that corresponds to a confidence score > 4 for the active sites (24% predicted FPR), ZiFOpT predicts that 15,565 (53%) of all zebrafish transcripts can be targeted *successfully* using OPEN. By restricting targets to those identified by ZiFOpT at a higher confidence score (> 8), the number of potential target sites for experimental testing could be reduced from 114,392 to 10,515, i.e., by ~ 90%. Thus, for

functional genomic studies, ZiFOpT is a valuable tool for identifying sites most amenable to targeting by ZFNs. Indeed, we have used ZiFOpT to predict activity for all 315,186 OPEN ZFN targets previously identified in zebrafish [8]. These results are presented in Supplemental Tables 3-27.

Table 2.3 | Summary of zebrafish OPEN ZFN target sites, classified by ZiFOpT

Confidence Score (Active Sites)	**False Positive Rate[1] (FPR)**	**# of zebrafish transcripts targeted[2]**		**Average # of ZFN target sites[2] in transcripts containing nuclease sites**	**# of potential target sites[2] eliminated by using ZiFOpT**	
**	**	25,174	(86%)	4.5	0	(0%)
> 4	24%	15,565	(53%)	2.3	78,934	(69%)
> 6	14%	12,622	(43%)	2.0	89,580	(78%)
> 8	7%	6,942	(24%)	1.5	103,877	(90%)

[1]estimated from training data [2]in coding exons 1-3 **no classification

DISCUSSION

Detailed analyses of available high resolution structures for DNA-protein complexes support the conclusion that there is no simple general code for DNA-protein recognition [36]. For certain classes of DNA binding proteins, including the C2H2 zinc finger proteins, it may be possible to decipher some of the rules that govern protein-DNA recognition by exploiting the increasing availability of data regarding sequence determinants of binding affinity and specificity. For example, Stormo's group has utilized contact propensities and weight matrices to predict which target sites a zinc finger motif is most likely to bind [27, 37]. Recently, Singh and colleagues utilized SVMs to predict whether a specific zinc finger protein will bind a specified target site [38]. Methods such as these utilize binding information for specific ZFPs interacting with a limited number of DNA target sites. In contrast, DNA microarray based experiments provide binding preferences of a transcription factor for thousands of potential sites [39-42]. These experiments should provide additional data for predicting and assessing transcription factor binding site models, including those for zinc finger proteins.

In the current study, we propose an approach for predicting whether a ZFP can be engineered to bind a specific DNA sequence without *a priori* knowledge of the ZFP amino

acid sequence. We analyzed base composition features and position-specific base propensities in a dataset of 135 different DNA target sites for which the OPEN selection method had been experimentally attempted. Our goal was to use this information to develop a rapid and reliable machine learning classifier to identify DNA sequences most amenable to site-specific targeting by zinc finger arrays generated using the OPEN design procedure. Based on our results, we developed a server-based application, ZiFOpT, which implements a sequence identity-based Naïve Bayes classifier, and identifies active OPEN target sites with an estimated average accuracy of 87%, specificity$^{+} \geq 90\%$ and sensitivity$^{+} \geq 94\%$ when evaluated using cross-validation and optimized for correlation coefficient. ZiFOpT performance on an independent "blind" test set of 66 experimentally validated ZFP targets was comparable, indicating that the performance measures are not likely biased by over-fitting.

In our statistical analysis of active versus inactive target sites, we detected biases in position-specific base composition of ZF targets (Figure 2.1). Thus, we anticipated that classifiers in which we attempted to capture base count biases or position-specific base propensities in the sequence encoding might perform as well as those using sequence identity, particularly in light of the size of the dataset relative to the size of the feature space for the sequence identity representation. For the Naïve Bayes classifier, however, sequence identity outperformed positional base counts and gave the best overall performance, in terms of the AUC of the ROC curve (0.89). For the SVM classifier, using positional base counts as input did provide substantially better performance than sequence identity (84% vs. 76%). Because the dataset used to train the SVM classifiers was smaller (to ensure a balanced number of positive and negative instances, see Methods), this difference in performance may be partly attributable to relatively sparse data for the sequence identity encoding.

Although the OPEN procedure tests only a small fraction of the total theoretical protein sequence space for the zinc finger recognition helix, it generates up to approximately 1 million ZFP combinations, clustered in what are expected to correspond to regions of optimal amino acid sequence space for the DNA target site of interest. Together with the results summarized in Figure 2.1, this suggests there are utilizable constraints on the DNA sequence space for 9-bp target sites that can be successfully targeted by ZFPs engineered by

OPEN. For example, the results in Figures 2.1 B and 2.1 C indicate that increased thymine content in target sites, especially at positions 1 and 7, may preclude high affinity or high specificity binding. Previous studies have suggested that ZFP recognition sites with relatively high purine nucleotide content are more often active targets for engineered zinc finger proteins [24, 29]. These earlier conclusions were based on analysis of target sites containing predominantly GNN-triplets and for ZFPs generated using modular assembly. The current analysis confirms and quantifies the contributions of high purine content as an important determinant of success for sequences targeted using OPEN. More specifically, our analyses indicate that for three-finger ZFPs, it is advisable to avoid target sites containing many thymine bases.

Based on the results reported here, ZiFOpT will be valuable for guiding investigators using OPEN to ZFN target sites with the greatest opportunities for success. The calculations shown in Table 3 illustrate the potential reduction in experimental effort that could be achieved by using ZiFOpT to identify ZFP target sites for every protein encoded by the zebrafish genome. Also, ZiFOpT should be valuable for selecting targets among the 695,819 total OPEN nuclease targets identified in protein-encoding transcripts of the human genome (Ensemble V51.1) [D. Reyon and J. Sander, unpublished], and could assist investigators who wish to apply OPEN technology to target specific genes or genomic regions of interest in other organisms. ZiFOpT classifies potential target sites for OPEN-generated ZFPs as "active" or "inactive" and provides a confidence score for the prediction. ZiFOpT is freely available and incorporated in the Zinc Finger Targeter (ZiFiT 3.2) web server (http://bindr.gdcb.iastate.edu/ZiFiT)[43]. ZiFiT can scan a given DNA sequence of interest and identify every potential DNA site targetable by OPEN. With the integration of ZiFOpT, users will be able to evaluate the expected success rate of OPEN for target sites identified by ZiFiT.

CONCLUSION

In this study, we developed machine learning classifiers that reliably identify DNA sites highly amenable to targeting by the OPEN zinc finger protein engineering method. Analysis of a dataset of 135 experimentally validated ZFP binding sites identified high

thymine content as a significant barrier to effective targeting by OPEN. In addition, comparison of results obtained using three different target sequence encodings as input for Naïve Bayes and SVM classifiers suggested that positional context plays a significant role in ZFP target site recognition. Importantly, however, a simple encoding based on sequence identity is sufficient to identify the most promising ZFP target sites, with ~87% accuracy. As more ZFP functional data become available and we learn more about the sequence composition of fingers in OPEN pools, our predictions should improve. At present, the ZiFOpT classifier presented here is expected to reduce the experimental effort required to identify an active ZFP-target site pair by ~75%, compared with selection of target sites without classification. By restricting experimental targets to "active" OPEN sites predicted with highest confidence, experimental success rates should be significantly enhanced. This in turn should accelerate the application of zinc finger proteins as tools for precise genetic manipulation in basic genomics research as well as in gene therapy.

METHODS

Definition of active and inactive ZFP target sites based on B2H assays

An *active* target site is a 9-bp DNA sequence for which the OPEN procedure has been used successfully to obtain at least one ZFP capable of binding the site with sufficient affinity and specificity to provide three-fold activation in a bacterial 2-hybrid (B2H) assay, i.e., to induce production of β-galactosidase by at least three-fold above the basal level of induction obtained using control constructs that lack the cognate ZFP target site [7, 28, 44]. An *inactive* target site is a 9-bp DNA sequence for which none of the corresponding OPEN-generated ZFPs tested were capable of producing a three-fold activation in the B2H assay.

Datasets of experimentally validated ZFP-target sites

ZFTS135 (cross-validation dataset): A zinc finger target site dataset generated from a group of 135 potential 9-bp zinc finger target sites (ZFTSs) that have been experimentally targeted using OPEN. For each ZFTS in the dataset, ZFPs have been selected using OPEN [7] and evaluated for DNA-binding activity *in vivo* using the B2H assay [9, 28, 44]. The sequences of all 135 ZFTS, together with their experimentally determined functional activity

labels (active or inactive) are provided in Supplemental Table 1. For 83 target sites in ZFTS135, functional activity labels, based on B2H assays, are reported here for the first time. The remaining 53 target sites, denoted by asterisks (*) were characterized previously [7, 28, 44] and experimental activity data were extracted from the Zinc Finger Database, ZiFDB (http://bindr.gdcb.iastate.edu:8080/ZiFDB/) [45].

ZFTS66 (independent test set): This dataset is an independent group of 66 potential 9-bp ZFN target sites (none of which overlap with those in ZFTS135), that have been experimentally targeted using OPEN. These sites were chosen by experts in the field, unaided by our classifier, in order to generate a "blind" test set for rigorous evaluation of ZiFOpT performance. 57 of these sites were determined to be 'active' based on B2H assay results, as described above. The sequences of all 66 ZFTS, along with classification and confidence scores, are provided in Supplemental Table 2.

Machine learning classifiers

Naïve Bayes is a probabilistic classifier that assumes the independence of each attribute and generates models that are amenable to user interpretation, usually without compromising performance [46]. We used the implementation available in the WEKA package version 3.5.7 [47]. For each instance, the classifier returns a classification of either "active" or "inactive" based on the posterior probability (Bayes' rule). The value of the classification threshold (θ) can be selected based on the desired trade-off between sensitivity and specificity. We evaluated several classification performance measures (see below), using a standard leave-one-out cross validation procedure.

Support Vector Machines (SVMs) find a hyper plane in high-dimensional space that maximizes the distance between the different classes of data in that space. We implemented the SVM classifier using the wrapper class available for LIBSVM [48]. We tested several different kernel functions. Best results were obtained using the radial basis function (RBF) kernel. Optimal cost and gamma parameters were determined using a grid search algorithm. Because SVM classifiers are sensitive to the number of positive and negative instances in the training set, and because our dataset is unbalanced (106 positive and 29 negative instances), we used a variation of the standard leave-one-out cross validation technique. For each test

case, we removed that instance and generated 10 randomized balanced training sets. The probability assigned to each test case was an average of the probability estimate generated from 10 randomized balanced training sets.

Target site sequence encoding

For each classifier, three different input sequence encodings were evaluated. The *sequence identity* input window consists of a target site represented as a 9 nucleotide DNA sequence, reading in the 5' to 3' direction on one strand (e.g., GTTGACGGC). The *base counts* input window consists of four single-digit values that represent the number of occurrences of each of the four DNA bases (G, A, C, T) within a target site (e.g., 4,1,2,2 for the target site in the preceding example). The *positional base counts* input window consists of a string of 12 values (3 sets of 4 digits), ranging from 0 to 3 and representing the number of times each base occurs in the first, second, and third positions within a triplet (e.g., 3,0,0,0;1,1,0,1;0,0,2,1, for the target site in the preceding example, in which G occurs in the first position of a triplet 3 times, once in the second and 0 times in the third.).

Classification performance measures

We used several standard performance measures: *accuracy, correlation coefficient (CC), specificity*$^+$, and *sensitivity*$^+$, and the AUC for standard ROC curves as described by Baldi et al. [49]. Here *True Positives* (TP) is the number of validated targets correctly predicted to be "active" target sites, i.e., sites that have been targeted successfully by an OPEN-generated ZFP to produce >3-fold activation in the B2H assay; *False Positives* (FP) is the number of "inactive" target sites incorrectly predicted to be "active" sites; *True Negative*s (TN) is the number of "inactive" target sites correctly predicted as such; *False Negatives* (FN) is the number of "active" target sites incorrectly predicted to be "inactive" sites.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad CC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

$$Specificity^+ = \frac{TP}{TP + FP} \qquad Sensitivity^+ = \frac{TP}{TP + FN}$$

$$\textit{False Positive Rate (FPR)} = \frac{FP}{FP + TN} \qquad \textit{True Positive Rate (FPR)} = \frac{TP}{TP + FN}$$

A *Receiver Operating Characteristic (ROC) curve* displays the tradeoff between the true positive rate (hit rate) and the false positive rate (false alarm rate) for different discrimination thresholds [49]. The Area Under the Curve (AUC) of the ROC plot is valuable for comparing performance of different classifiers because it portrays the tradeoff between the false positive rate and the true positive over the range of classification threshold values.

Confidence Score

The posterior probability returned by ZiFOpT for classifying each target site was used to generate a confidence score. Target sites with posterior probability were classified ‘active’ if they had posterior probability ≥ 0.5 and ‘inactive’ otherwise. For the ‘active’ class, the posterior probability was transformed to a scale from 0 to 9 by incrementing the confidence score by 1 as the posterior probability increased by 0.05 above 0.5. Therefore, a posterior probability of 0.75 corresponds to an ‘active’ classification with a confidence score of 5. For the ‘inactive’ class, the confidence score was incremented by 1 as the posterior probability decreased by 0.05 below 0.5. Therefore, a posterior probability of 0.25 corresponds to an ‘inactive’ classification with a confidence of 5.

COMPETING INTERESTS

J.K.J. is an inventor on a patent application describing the OPEN method. The remaining author(s) declare that they have no competing interests.

AUTHORS’ CONTRIBUTIONS

JS was responsible for experimental design, analysis of results, initial draft of manuscript, participated in discussions and manuscript revisions. DR parsed the data, ran the machine learning algorithms, participated in discussions and manuscript revisions. MM, ST, JF, and JK generated the experimental data and participated in manuscript reviews. FF, DD,

DR, and DV participated in discussions, analysis of results, and manuscript revisions. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

This work was supported in part by the following grants: NIH T32CA009216 (J.D.S); NIH GM066387 (D.D.); NIH GM069906 and GM078369 (J.K.J.); NSF DBI 0501678 (D.F.V.) and graduate research assistantships provided by USDA MGET 2001-52100-11506, NSF IGERT0504304, and ISU's Center for Integrated Animal Genomics (CIAG) and Department of Genetics, Development and Cell Biology. We thank members of our groups, especially M. Terribilini, B. Lewis, and P. Zaback for valuable comments.

REFERENCES

1. Cathomen T, Joung JK: **Zinc-finger nucleases: the next generation emerges**. *Mol Ther* 2008, **16**(7):1200-1207.

2. Carroll D: **Progress and prospects: zinc-finger nucleases as gene therapy agents**. *Gene Ther* 2008, **15**(22):1463-1468.

3. Morton J, Davis MW, Jorgensen EM, Carroll D: **Induction and repair of zinc-finger nuclease-targeted double-strand breaks in Caenorhabditis elegans somatic cells**. *Proc Natl Acad Sci U S A* 2006, **103**(44):16370-16375.

4. Santiago Y, Chan E, Liu PQ, Orlando S, Zhang L, Urnov FD, Holmes MC, Guschin D, Waite A, Miller JC *et al*: **Targeted gene knockout in mammalian cells by using engineered zinc-finger nucleases**. *Proc Natl Acad Sci U S A* 2008, **105**(15):5809-5814.

5. Beumer K, Bhattacharyya G, Bibikova M, Trautman JK, Carroll D: **Efficient gene targeting in Drosophila with zinc-finger nucleases**. *Genetics* 2006, **172**(4):2391-2403.

6. Doyon Y, McCammon JM, Miller JC, Faraji F, Ngo C, Katibah GE, Amora R, Hocking TD, Zhang L, Rebar EJ *et al*: **Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases**. *Nat Biotechnol* 2008, **26**(6):702-708.

7. Maeder ML, Thibodeau-Beganny S, Osiak A, Wright DA, Anthony RM, Eichtinger M, Jiang T, Foley JE, Winfrey RJ, Townsend JA *et al*: **Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification**. *Mol Cell* 2008, **31**(2):294-301.

8. Foley JE, Yeh JR, Maeder ML, Reyon D, Sander JD, Peterson RT, Joung JK: **Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN)**. *PLoS ONE* 2009, **4**(2):e4348.

9. Townsend JA, Wright DA, Winfrey RJ, Fu F, Maeder M, Joung JK, Voytas DF: **High-frequency modification of plant genes using engineered zinc-finger nucleases**. *Nature* 2009, **In Press doi 10.103/nature07845**.

10. Lee HJ, Kim E, Kim JS: **Targeted chromosomal deletions in human cells using zinc finger nucleases**. *Genome Res* 2009, **20**(1):81-89.

11. Shukla VK, Doyon Y, Miller JC, DeKelver RC, Moehle EA, Worden SE, Mitchell JC, Arnold NL, Gopalan S, Meng X *et al*: **Precise genome modification in the crop species Zea mays using zinc-finger nucleases**. *Nature* 2009, **459**(7245):437-441.

12. Voigt B, Serikawa T: **Pluripotent stem cells and other technologies will eventually open the door for straightforward gene targeting in the rat**. *Dis Model Mech* 2009, **2**(7-8):341-343.

13. Geurts AM, Cost GJ, Freyvert Y, Zeitler B, Miller JC, Choi VM, Jenkins SS, Wood A, Cui X, Meng X *et al*: **Knockout rats via embryo microinjection of zinc-finger nucleases**. *Science* 2009, **325**(5939):433.

14. Zou J, Maeder ML, Mali P, Pruett-Miller SM, Thibodeau-Beganny S, Chou BK, Chen G, Ye Z, Park IH, Daley GQ *et al*: **Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells**. *Cell Stem Cell* 2009, **5**(1):97-110.

15. Scott CT: **The zinc finger nuclease monopoly**. *Nat Biotechnol* 2005, **23**(8):915-918.

16. Kaiser J: **Gene therapy. Putting the fingers on gene repair**. *Science* 2005, **310**(5756):1894-1896.

17. Pearson H: **Protein engineering: The fate of fingers**. *Nature* 2008, **455**(7210):160-164.

18. Klug A: **Towards therapeutic applications of engineered zinc finger proteins**. *FEBS Lett* 2005, **579**(4):892-894.

19. Blancafort P, Segal DJ, Barbas CF, 3rd: **Designing transcription factor architectures for drug discovery**. *Mol Pharmacol* 2004, **66**(6):1361-1371.

20. Bae KH, Kwon YD, Shin HC, Hwang MS, Ryu EH, Park KS, Yang HY, Lee DK, Lee Y, Park J *et al*: **Human zinc fingers as building blocks in the construction of artificial transcription factors**. *Nat Biotechnol* 2003, **21**(3):275-280.

21. Liu Q, Xia Z, Zhong X, Case CC: **Validated zinc finger protein designs for all 16 GNN DNA triplet targets**. *J Biol Chem* 2002, **277**(6):3850-3856.

22. Segal DJ, Dreier B, Beerli RR, Barbas CF, 3rd: **Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences**. *Proc Natl Acad Sci U S A* 1999, **96**(6):2758-2763.

23. Segal DJ, Beerli RR, Blancafort P, Dreier B, Effertz K, Huber A, Koksch B, Lund CV, Magnenat L, Valente D *et al*: **Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins**. *Biochemistry* 2003, **42**(7):2137-2148.

24. Ramirez CL, Foley JE, Wright DA, Muller-Lerch F, Rahman SH, Cornu TI, Winfrey RJ, Sander JD, Fu F, Townsend JA *et al*: **Unexpected failure rates for modular assembly of engineered zinc fingers**. *Nat Methods* 2008, **5**(5):374-375.

25. Sander JD, Zaback P, Joung JK, Voytas DF, Dobbs D: **An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins**. *Nucleic Acids Res* 2009, **37**(2):506-515.

26. Kim HJ, Lee HJ, Kim H, Cho SW, Kim JS: **Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly**. *Genome Res* 2009, **19**(7):1279-1288.

27. Liu J, Stormo GD: **Context-dependent DNA recognition code for C2H2 zinc-finger transcription factors**. *Bioinformatics* 2008, **24**(17):1850-1857.

28. Hurt JA, Thibodeau SA, Hirsh AS, Pabo CO, Joung JK: **Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection**. *Proc Natl Acad Sci U S A* 2003, **100**(21):12271-12276.

29. Meng X, Noyes MB, Zhu LJ, Lawson ND, Wolfe SA: **Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases**. *Nat Biotechnol* 2008, **26**(6):695-701.

30. Carroll D, Morton JJ, Beumer KJ, Segal DJ: **Design, construction and in vitro testing of zinc finger nucleases**. *Nat Protoc* 2006, **1**(3):1329-1341.

31. Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K: **Improved visualization of protein consensus sequences by iceLogo**. *Nat Methods* 2009, **6**(11):786-787.

32. Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D: **Prediction of RNA binding sites in proteins from amino acid sequence**. *Rna* 2006, **12**(8):1450-1462.

33. Narlikar L, Hartemink AJ: **Sequence features of DNA binding sites reveal structural class of associated transcription factor**. *Bioinformatics* 2006, **22**(2):157-163.

34. Capra JA, Singh M: **Predicting functionally important residues from sequence conservation**. *Bioinformatics* 2007, **23**(15):1875-1882.

35. Punta M, Ofran Y: **The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function**. *PLoS Comput Biol* 2008, **4**(10):e1000160.

36. Pabo CO, Nekludova L: **Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition?** *J Mol Biol* 2000, **301**(3):597-624.

37. Benos PV, Lapedes AS, Stormo GD: **Probabilistic code for DNA recognition by proteins of the EGR family**. *J Mol Biol* 2002, **323**(4):701-727.

38. Persikov AV, Osada R, Singh M: **Predicting DNA recognition by Cys2His2 zinc finger proteins**. *Bioinformatics* 2009, **25**(1):22-29.

39. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, Young RA, Bulyk ML: **Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays**. *Nat Genet* 2004, **36**(12):1331-1339.

40. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML: **Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities**. *Nat Biotechnol* 2006, **24**(11):1429-1435.

41. Ragoussis J, Field S, Udalova IA: **Quantitative profiling of protein-DNA binding on microarrays**. *Methods Mol Biol* 2006, **338**:261-280.

42. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, Alleyne TM, Mnaimneh S, Botvinnik OB, Chan ET *et al*: **Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences**. *Cell* 2008, **133**(7):1266-1276.

43. Sander JD, Zaback P, Joung JK, Voytas DF, Dobbs D: **Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W599-605.

44. Joung JK, Ramm EI, Pabo CO: **A bacterial two-hybrid selection system for studying protein-DNA and protein-protein interactions**. *Proc Natl Acad Sci U S A* 2000, **97**(13):7382-7387.

45. Fu F, Sander JD, Maeder M, Thibodeau-Beganny S, Joung JK, Dobbs D, Miller L, Voytas DF: **Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays**. *Nucleic Acids Res* 2009, **37**(Database issue):D279-283.

46. Buntine W: **Theory refinement on Bayesian networks**. In: *Proceedings of the seventh conference (1991) on Uncertainty in artificial intelligence.* Los Angeles, California, United States: Morgan Kaufmann Publishers Inc.; 1991.

47. Witten IH, Frank E: **Data mining: practical machine learning tools and techniques**, 2nd edn. San Francisco: Morgan Kaufman; 2005.

48. Chang CC, Lin CJ: **LIBSVM: a library for support vector machines**. 2001.

49. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H: **Assessing the accuracy of prediction algorithms for classification: an overview**. *Bioinformatics* 2000, **16**(5):412-424.

Supplemental Table 2.1. ZFTS135 dataset of zinc finger target sequences and activity labels. DNA sequences of zinc finger target sites (shown reading 5' - 3') and functional activity labels for 135 experimentally validated ZFP-target site pairs analyzed in this study. "Active" indicates sites for which the OPEN method has been used to generate at least one corresponding zinc finger array that provides ≥ 3-fold activation of a β-galactosidase reporter gene in a bacterial 2-hybrid (B2H) reporter assay. For 82 target sites, B2H assay results are reported here for the first time. Asterisks (*) denote the remaining 53 target sites for which B2H activity results were reported previously and were extracted from ZiFDB (http://bindr.gdcb.iastate.edu/ZiFDB/).

Target 5' - 3'	Activity Label
GCATTTTTT	Inactive
GCCGCAGCT	Inactive
GCTTGAGCT	Inactive
*GGCGCCTAC	Inactive
*GGCGGAGAT	Inactive
*GGCGTTTGC	Inactive
GGGGAATAC	Inactive
GGTGTCTTT	Inactive
*GTGGCGGAT	Inactive
GTGTTTGAA	Inactive
GTTGATTTT	Inactive
GTTTTTGAG	Inactive
TAAGTTGTT	Inactive
TCTGCATTC	Inactive
*TCTGGCGCT	Inactive
*TCTGGTTTC	Inactive
TCTGTGTTC	Inactive
TCTTGGGTA	Inactive
*TGCGGCTGT	Inactive
TGCTGAGAC	Inactive
TGCTTTGTT	Inactive
TGGGTAGAA	Inactive
TGGTTTGTA	Inactive
TGTGTGTTC	Inactive
TTAGGAGGT	Inactive
TTTGAGGAT	Inactive
TTTGCTGAA	Inactive
TTTGTGGTG	Inactive
TTTGTGTGT	Inactive
*GAAGAAGCT	Active
*GAAGACGCT	Active
*GAAGATGGT	Active
*GAAGCAGCA	Active
*GAAGGATTC	Active
*GAAGTGGTC	Active
GAATTGGCG	Active
*GACGACGGC	Active
GACGCCGGA	Active
*GACGCTGCT	Active
GACTGAGAA	Active
GACTGGGCG	Active

Target 5' - 3'	Activity Label
GAGGGCGGC	Active
GAGGGGGCG	Active
*GAGTGAGGA	Active
GAGTGAGTT	Active
*GAGTTTGCC	Active
GATGAAGAC	Active
*GATGAAGCT	Active
GATGAGGCG	Active
GATGCCAAC	Active
GATGCGGCA	Active
GATGTAGCC	Active
GATTGAGTT	Active
GCAGAAGCT	Active
*GCAGCAGAG	Active
*GCAGCAGGA	Active
*GCAGCGGGC	Active
*GCAGGAGGT	Active
GCAGTGTGT	Active
GCCGAAGAT	Active
*GCCGCCGGC	Active
GCCGCTGGA	Active
*GCCGCTGGG	Active
*GCCGGCGGC	Active
GCCGGTGCA	Active
*GCCGGTGGC	Active
*GCCGTCGCC	Active
GCGGCCGCG	Active
*GCGGCGGAC	Active
*GCGGCTGGG	Active
GCGGGGGGC	Active
GCGGGTGTG	Active
GCGGTGGCG	Active
GCGTGGGCG	Active
GCGTTGGCG	Active
GCTGACTTT	Active
GCTGAGGCT	Active
*GCTGATGCC	Active
GCTGCAGAA	Active
GCTGCCGTC	Active
*GCTGCTGCC	Active
GCTGCTGGT	Active

Target 5' - 3'	Activity Label
*GGAGGGGCT	Active
GGAGGTGAC	Active
GGAGGTGGA	Active
*GGAGGTGGT	Active
GGCGCAAAC	Active
*GGCGGCGGA	Active
GGCTGAGGC	Active
GGCTGGGCT	Active
GGCTGGGTG	Active
*GGGGAAGAG	Active
GGGGAAGAT	Active
*GGGGACGTC	Active
*GGGGAGGAG	Active
GGGGTCGAC	Active
GGGGTGGGT	Active
GGTGAATTT	Active
GGTGAGGCA	Active
GGTGCAGCA	Active
GGTGCTGAC	Active
GGTGGCGCT	Active
GGTGGGGTG	Active
GTAGAGGAG	Active
*GTAGATGGA	Active
GTAGCCTGT	Active
GTAGCTGGA	Active
GTCGACGCC	Active
*GTCGATGCC	Active
*GTCGGGGTA	Active
GTCTGAGGC	Active
GTCTGGGCT	Active
*GTGGACGCG	Active
*GTGGCTGGT	Active
*GTGTAGGGG	Active
TAAGCAGAA	Active
TAATTGGAG	Active
TCTGCTGGC	Active
TCTGGTGAG	Active
*TGGGAGTCT	Active
TGGGATGTT	Active
*TGGGGTGCC	Active
*TGGGTGGCA	Active

Target 5' - 3'	Activity Label
GACTGGGCT	Active
*GAGGACGGC	Active
*GAGGACGTG	Active
*TGGGCTGCT	Active

Target 5' - 3'	Activity Label
GCTGGAGGG	Active
GCAGCGGGA	Active
*GGAGGAGGT	Active
GGAGGCGTG	Active

Target 5' - 3'	Activity Label
*TTAGAAGTG	Active
*TTATGGGAG	Active
TTTGTTGGC	Active
*GGTGCTGCC	Active

Supplemental Table 2.2. ZFTS66 dataset of 66 experimentally validated zinc finger target site sequences, used as an independent ("blind") test set in this study. Target 5' - 3' = DNA sequence of zinc finger target site; Activity Label = actual activity, determined experimentally; Prediction: activity predicted by ZiFOpT; Confidence Score = confidence in prediction, ranging from 0 (lowest) to 9 (highest); see Methods for additional details.

Target 5' - 3'	Activity Label	Prediction	Confidence Score
GGTGGAGCA	Active	Active	9
GGTGTCGAA	Active	Active	9
GTAGAAGAG	Active	Active	9
GTAGCAGTC	Active	Active	9
GTAGCTGCG	Active	Active	9
GTCGTTGCC	Active	Active	9
GTCTGAGTA	Active	Active	9
GTGGATGGT	Active	Active	9
GTGGCAGGA	Active	Active	9
GTGGCCGTG	Active	Active	9
TAATTGGGG	Active	Active	9
TCTGAGGAC	Active	Active	9
TCTGGTGAC	Active	Inactive	9
TGGGATGTG	Active	Active	9
TGGGCAGTG	Active	Active	9
TGGGGGGCA	Active	Active	9
TGGGTCGAC	Active	Inactive	9
TGTGACGGC	Active	Active	9
TGTGGGGGG	Active	Active	9
TTAGGGGAC	Active	Active	9
TGGGATGGA	Active	Active	9
GACGGCAAC	Active	Active	9
GTAGAGGGT	Active	Active	9
GCCGGAGAC	Active	Active	9
GCATGGGCA	Active	Active	9
GGTGATGCT	Active	Active	9
GCCGAAGAG	Active	Active	9
GACGGCTGT	Active	Active	9
GCTGCAGGT	Active	Active	9
GAGGATGTA	Active	Active	9
GCCGAAGTT	Active	Active	9
GACGGAGCT	Active	Active	9
GCTGATGGC	Active	Active	9
GCGGTTGCA	Active	Active	9
GACGGAGTC	Active	Active	9
GCAGGTGGA	Active	Active	9

Target 5' - 3'	Activity Label	Prediction	Confidence Score
GGGGAAGGT	Active	Active	9
GCCGCAGTG	Active	Active	9
GATGGTGAG	Active	Active	9
GGTTGGGAG	Active	Active	9
GCAGGCGCA	Active	Active	9
GAGGAGGGT	Active	Active	9
GGGGAAGGA	Active	Active	9
GAGGAGAAC	Active	Active	9
GGAGCCGGC	Active	Active	9
GCTGAGGGG	Active	Active	9
GCAGAAGTA	Active	Active	9
GAAGTAGCA	Active	Active	9
GCTGAAGCG	Active	Active	9
GATGATGGC	Active	Active	9
GAGGAAGCT	Active	Active	9
GTGGATGCA	Active	Active	9
GTGGCAGAA	Active	Active	9
TAAGAAGAG	Active	Active	9
GACGGAGGA	Active	Active	9
GATGAAGAA	Active	Active	9
GTAGCGGGT	Active	Active	9
GGTTAGGAT	Active	Active	9
GCGGCGGCC	Active	Active	9
GGTTGAGCG	Active	Active	9
GAGGAGGAG	Active	Active	9
GAGGCGTGT	Active	Active	9
GGAGGTGAG	Active	Active	9
GGAGGTGCC	Active	Active	9
GAAGAAGAG	Active	Active	9
GCGGCCGAA	Active	Active	9
GGAGAAGTA	Active	Active	9
GCTGAGGGC	Active	Active	9
GAGGACTGC	Active	Active	9
GGGGCTGCA	Active	Active	9
GAGGTAGTG	Active	Active	9
GAGGCGGAC	Active	Active	9
TGCGATGGA	Active	Active	9
GCTGGTGTC	Active	Active	9
TGGGCCGAC	Active	Active	9
GAGGCAGAA	Active	Active	9
GAAGCAGGC	Active	Active	9
GAGGATGGG	Active	Active	9
GCATGAGCT	Active	Active	9

Target 5' - 3'	Activity Label	Prediction	Confidence Score
GCTGGTGGC	Active	Active	9
GAGGCCTGT	Active	Active	9
GCTGCGGTG	Active	Active	9
GGAGGAGAT	Active	Active	9
GTGGTGGCT	Active	Active	9
GGATGAGCC	Active	Active	9
GCTGACTGC	Active	Active	9
GCGGGAGGG	Active	Active	8
GCGGTAGCT	Active	Active	8
GCTGACGGT	Active	Active	8
GCTGAGGAA	Active	Active	8
GCTGCAGAA	Active	Active	8
GCTGGTGAA	Active	Active	8
GCTGTCGAA	Active	Active	8
GCTGTTGGG	Active	Active	8
GGAGACGGT	Active	Active	8
GGCGACGGC	Active	Active	8
GGCGAGGAA	Active	Active	8
GGCGCAGGG	Active	Active	8
GGGGCAGTG	Active	Active	8
GGGGCGGGT	Active	Active	8
GGGGCTGAG	Active	Active	8
GGGGGAGGG	Active	Active	8
GGTGAAGAG	Active	Active	8
GGTGCCGAG	Active	Active	8
GACTTTGGT	Inactive	Active	7
GAGGCAGCA	Active	Active	7
GAGGCCGAG	Active	Active	7
GAGGCCGGC	Active	Active	7
GAGGGAGGA	Active	Active	7
GAGGTGGGT	Active	Active	7
GCAGCAGGG	Active	Active	7
GCAGGGGCG	Active	Active	7
GCAGGTGCT	Active	Active	7
GCCGCGGCC	Active	Active	7
GCGGCTGCC	Active	Active	7
GCGGCTGCG	Active	Active	7
GAAGGGTGC	Active	Active	6
GAAGGTGTT	Active	Active	6
GAAGTCTGC	Active	Active	6
GACGAAGGC	Active	Active	6
GACGACGAA	Active	Active	6
GAGGAGGTC	Active	Active	6

Target 5' - 3'	Activity Label	Prediction	Confidence Score
GTCGTGGCC	Inactive	Active	5
GTAGGAGAG	Inactive	Active	5
GTCGGCGTA	Inactive	Active	5
GGTGCTGCG	Inactive	Active	5
GAAGGGGCC	Active	Active	5
GCAGCCGCA	Inactive	Active	4
GGAGTTGTT	Inactive	Active	4
GTCTGAGCA	Inactive	Active	4
GGGTTTGCA	Inactive	Active	4
GGTGATGAA	Inactive	Active	3
GTCGCAGTA	Inactive	Active	3
GCTTAGGGT	Inactive	Active	3
GCGTTTGAG	Inactive	Active	2
GTCGCTGTC	Inactive	Active	1
TCTGGAGAT	Inactive	Inactive	1
TGTGAATGT	Inactive	Inactive	1
GGCGGAGCA	Inactive	Active	1
GGTTTTGAG	Inactive	Active	0

CHAPTER 3. ZFNGENOME: A COMPREHENSIVE RESOURCE FOR LOCATING ZINC FINGER NUCLEASE TARGET SITES IN MODEL ORGANISMS

Deepak Reyon*, Jessica R. Kirkpatrick, Jeffry D. Sander, Feng Zhang, Daniel F. Voytas, J. Keith Joung, Drena Dobbs and Clark R. Coffman. ZFNGenome: A comprehensive resource for locating zinc finger nuclease target sites in model organisms. *BMC Genomics*. 2011 Jan. 28.

* Corresponding Author

ABSTRACT

Background

Zinc Finger Nucleases (ZFNs) have tremendous potential as tools to facilitate genomic modifications, such as precise gene knockouts or gene replacements by homologous recombination. ZFNs can be used to advance both basic research and clinical applications, including gene therapy. Recently, the ability to engineer ZFNs that target any desired genomic DNA sequence with high fidelity has improved significantly with the introduction of rapid, robust, and publicly available techniques for ZFN design such as the Oligomerized Pool ENgineering (OPEN) method. The motivation for this study is to make resources for genome modifications using OPEN-generated ZFNs more accessible to researchers by creating a user-friendly interface that identifies all potential ZFN target sites in the complete genomes of seven model organisms.

Description

ZFNGenome is a GBrowse-based tool for identifying and visualizing potential target sites for OPEN-generated ZFNs. ZFNGenome currently includes a total of more than 11.6 million potential ZFN target sites, mapped within the fully sequenced genomes of seven model organisms. These include: *S. cerevisiae, C. reinhardtii, A. thaliana, D. melanogaster, D. rerio, C. elegans*, and *H. sapiens;* additional model organisms will be included in future updates. ZFNGenome provides researchers with information about each potential ZFN target

site, including its chromosomal location, position relative to transcription initiation site(s), and frequency of occurrence within the genome. Users can query ZFNGenome using several different criteria (e.g., gene ID, transcript ID, or target site sequence). Targets identified using ZFNGenome can be visualized at multiple scales within the flexible GBrowse 1.7 environment and can be imported as annotations into other genome browsers. ZFNGenome is dynamically linked to ZiFDB, allowing users access to all available information about zinc finger reagents, such as the effectiveness of a given ZFN in creating double-stranded breaks.

Conclusions

ZFNGenome provides a user-friendly interface that allows researchers to access resources and information regarding genomic target sites for engineered ZFNs in seven model organisms. This genome-wide database of potential ZFN target sites should greatly facilitate the utilization of ZFNs in both basic and clinical research.

ZFNGenome is freely available at: *http://bindr.gdcb.iastate.edu/ZFNGenome*.

BACKGROUND

The ability to efficiently modify the genome of an organism with a high degree of specificity would advance both research with model organisms and human gene therapy [1-3]. In recent studies, zinc finger nuclease (ZFN)-mediated genomic modification rates of 3% - 100% for specific genes have been reported in zebrafish, *Arabidopsis*, and rat [4-16]. Moreover, ZFNs are being evaluated in human gene therapy clinical trials for treating AIDS [11, 17-19]. Thus, ZFNs are emerging as premier tools for site-specific genomic modification in both animals and plants.

Engineered ZFNs consist of two zinc finger arrays (ZFAs), each of which is fused to a single subunit of a non-specific endonuclease, such as the nuclease domain from the *Fok*I enzyme, which becomes active upon dimerization [20, 21]. Typically, a single ZFA consists of 3 or 4 zinc finger domains, each of which is designed to recognize a specific nucleotide triplet (GGC, GAT, etc.) [22]. Thus, ZFNs composed of two “3-finger” ZFAs are capable of recognizing an 18 base pair target site; an 18 base pair recognition sequence is generally unique, even within large genomes such as those of humans and plants. By directing the co-localization and dimerization of two FokI nuclease monomers, ZFNs generate a functional

site-specific endonuclease that creates a double-stranded break (DSB) in DNA at the targeted locus [23] (Figure. 3.1 A).

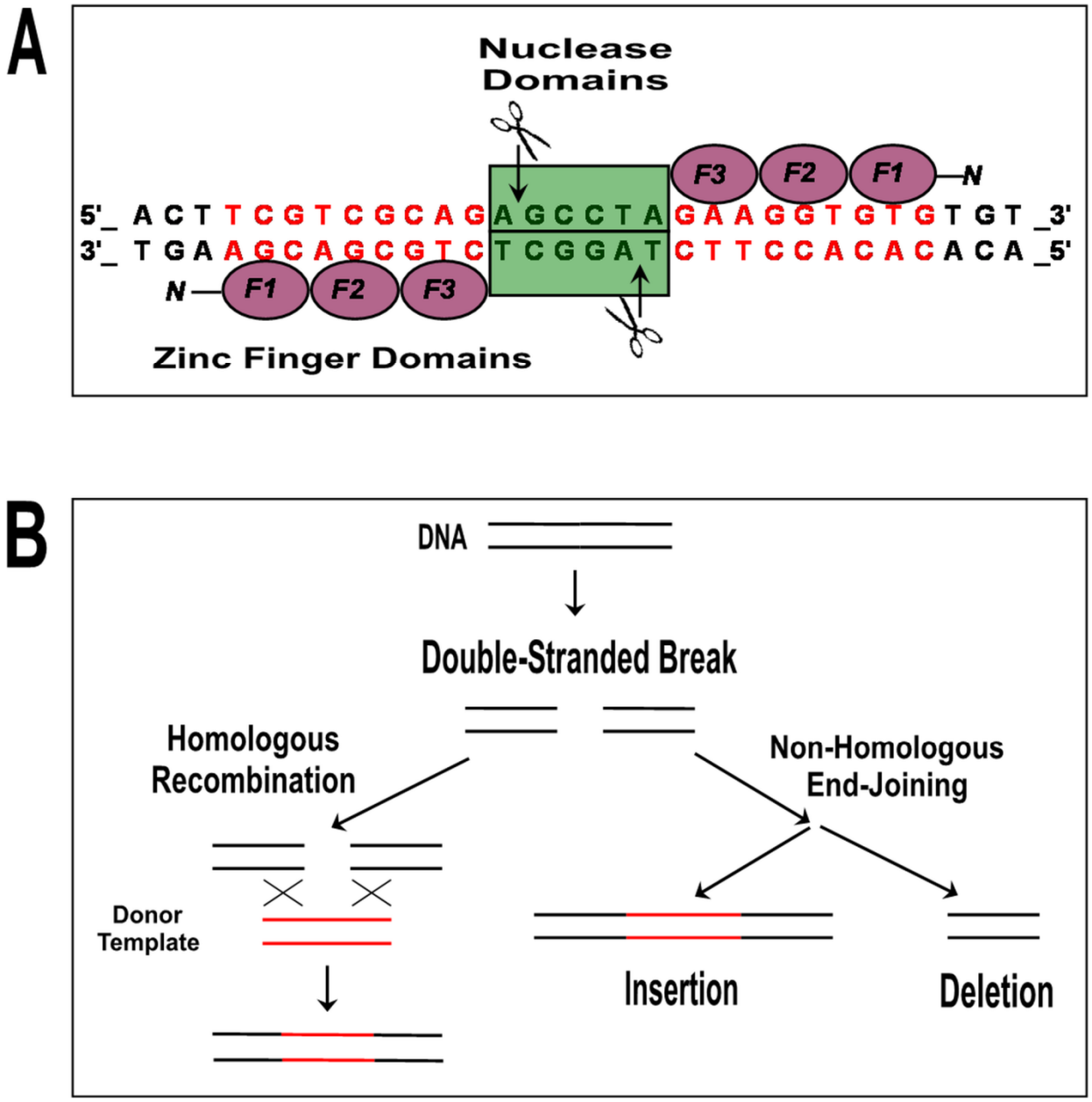


Figure 3.1. ZFNs generate site-specific double-stranded breaks that can be used for homologous recombination or mutagenesis. (A) ZFNs are composed of two arrays that recognize 9-12 base pairs each. Two arrays with three fingers, F1-F2-F3, that recognize nine base pairs each are shown. Each array is fused to one half of a nonspecific *Fok*I endonuclease (green). Upon dimerization, the *Fok*I endonuclease is activated and creates a double-stranded break at sites flanked by the DNA binding sites recognized by the zinc finger arrays. Scissors and arrows denote the cut sites. (B) In most cells, double-stranded breaks (DSBs) are repaired by one of two major pathways. If a donor template is

available, homologous recombination can result in engineered nucleotide substitutions at the target site (left). Alternatively, DSBs can be repaired by non-homologous end-joining, an error-prone mechanism that frequently results in small deletions or insertions at the site of the DSB (right).

In eukaryotes, repair of DSBs in DNA is primarily accomplished via one of two pathways, homologous recombination (HR) and non-homologous end-joining (NHEJ) (Figure 3.1 B)**.** Depending on the desired modification, either pathway can be exploited in ZFN-mediated genomic engineering. Because HR relies on homologous DNA to repair the DSB, gene targeting can be achieved by supplying an exogenous "donor" template. This results in replication of the "donor" DNA sequence at the target locus, a process that has been utilized to introduce small mutations or large insertions [4, 9, 12, 13, 16, 24-27]. In contrast, NHEJ is an error-prone repair process and hence is ideal for generating mutations that can result in gene knockouts or knock-downs when the ZFN-mediated DSB is introduced into the protein coding sequence of a gene [5-9, 11, 28, 29].

Oligomerized Pool Engineering (OPEN) is a highly robust and publicly available protocol for engineering zinc finger arrays with high specificity and *in vivo* functionality [9, 30, 31]. OPEN has been successfully used to generate ZFNs that function efficiently in plants [13, 15], zebrafish [6], and human somatic [9] and pluripotent stem cells [16]. OPEN is a selection-based method in which a pre-constructed randomized pool of candidate ZFAs is screened to identify those with high affinity and specificity for a desired target sequence. Significantly higher *in vivo* success rates have been reported using OPEN-generated ZFNs, compared with ZFNs generated using the more traditional modular assembly approach [32-34]. Resources for generating ZFNs using OPEN have been developed and made publicly available by the Zinc Finger Consortium [9, 31, 35]. Currently, OPEN reagents include modules that recognize all 16 possible GNN triplets (i.e., DNA triplets beginning with G, followed by any nucleotide in the second and third positions), as well as several TNN triplets. Thus, all DNA sites that contain only GNN and/or select TNN triplets can potentially be targeted using the OPEN protocol [9].

To facilitate use of OPEN ZFNs for genome modification, we have developed *ZFNGenome*, a resource that displays potential ZFN target sites in a genome browser built on the user-friendly GBrowse platform [36]. We analyzed the complete sequenced genomes of seven model organisms and identified all sequences that are potentially targetable using

currently available OPEN ZFN reagents. ZFN reagents were obtained from Joung and colleagues [9] and ZFN target sites were identified using software implemented in the ZiFiT web server [37, 38]. ZFNGenome thus allows users to quickly evaluate “pre-identified” ZFN target sites for any desired gene or region of interest.

To our knowledge, ZFNGenome represents the first compendium of potential ZFN target sites in sequenced and annotated genomes of model organisms. The current version includes ZFN target sites in seven organisms: *Saccharomyces cerevisiae* (budding yeast), *Chlamydomonas reinhardtii* (green algae), *Arabidopsis thaliana* (thale cress), *Caenorhabditis elegans* (nematode), *Drosophila melanogaster* (fruit fly), *Danio rerio* (zebrafish), and *Homo sapiens* (human). Additional model organisms, including three plant species; *Glycine max* (soybean), *Oryza sativa* (rice), *Zea mays* (maize), and three animal species *Tribolium castaneum* (red flour beetle), *Mus musculus* (mouse), *Rattus norvegicus* (brown rat) will be added in the near future.

CONSTRUCTION AND CONTENT

The motivation for implementing ZFNGenome, summarized in Figure 3.2, was to create a user-friendly interface between two valuable open-source genomic resources: i) established genome browsers, with associated genomic DNA sequences, annotations and other resources available for model organisms; and ii) ZFN design software tools and experimental reagents made available by the Zinc Finger Consortium. ZFNGenome integrates these resources by allowing users to visualize all potential ZFN target sites in a chosen gene or genomic region of a sequenced model organism, with flexible viewing

options and annotated genomic features provided in a GBrowse interface.

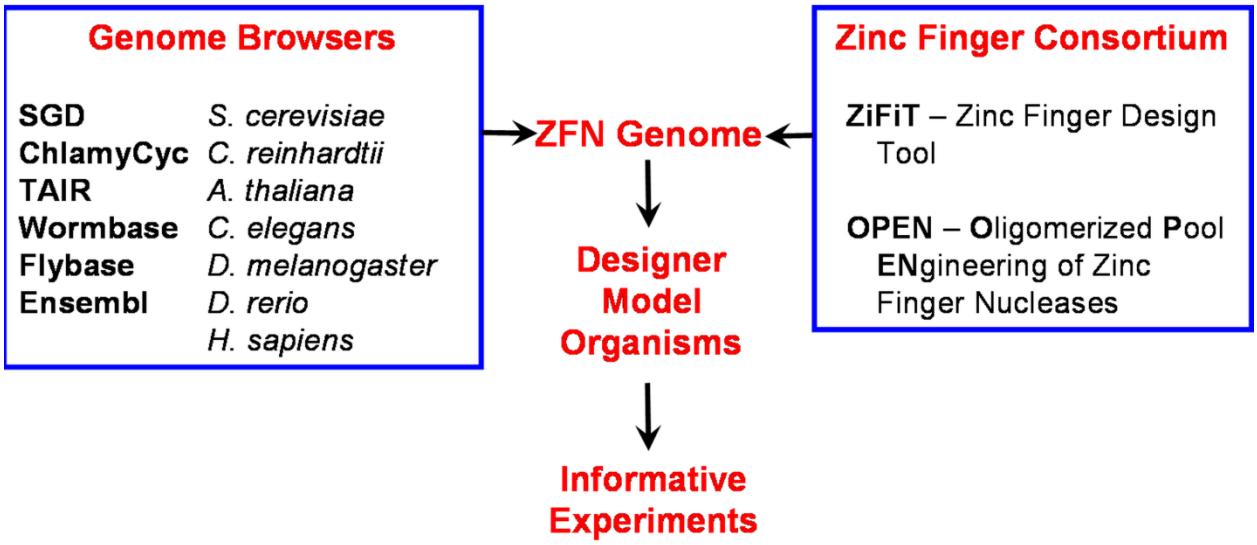


Figure 3.2. An overview of the ZFNGenome architecture. ZFNGenome creates a user-friendly interface for visualizing all potential ZFN target sites in seven model organisms by integrating the genomic information from genome browsers of sequenced and annotated genomes with the tools of the Zinc Finger Consortium. This interface allows researchers using these model organisms to easily determine whether ZFNs are available for the design and execution of targeted genome modifications.

Table 1 lists the organisms for which complete genomic sequence data were analyzed in this study, along with the data sources for genomic DNA sequences and annotations. The number of potential OPEN target sites identified is shown for each organism. To identify all potential OPEN ZFN target sites, annotated complete genome sequence files were scanned using the ZiFiT algorithm [38], which was modified to accommodate the sequences of an entire chromosome. Only sites for which ZFNs can be engineered using currently available OPEN reagents and spacer distances between the two ZFAs of 5, 6 or 7 base pairs were included [9]. Because OPEN selections are performed in a Dam+/Dcm+ *E. coli strain,* genomic target sequences that contain potential *dam* or *dcm* methylation sites were excluded from consideration, as were sites that lacked a GNN sub site. Previous studies have shown that most successful OPEN sites contain at least one GNN [31].

ZFNGenome utilizes GBrowse 1.7 [36] to display identified potential OPEN target sites, along with basic genome annotations, such as genes, transcripts, exons, introns, and 5' and 3' UTRs. ZFNGenome is hosted on an Apache2 web server and uses a MySQL DB

linked to a GBrowse front end via open source adaptors available in BioPerl (version 1.6) [39]. The ZFN target sites can be exported for use as annotations in other GBrowse-based genome browsers such TAIR and Wormbase. As described below, each ZFN target site is hyperlinked to ZiFDB [40].

Table 3.1. Model organism genomes analyzed and the number of OPEN ZFN target sites identified.

Organism	Source [1]	Total # of OPEN target sites[2]	Total # of transcripts[2]	ZFN targetable transcripts [2]		Avg. # ZFN target sites per transcript	GC Content
				#	%		
Saccharomyces	**SGD**	**31,822**	**6,685**	**5,810**	**87**	**5.5**	**38.3**
Chlamydomonas	**ChlamyCyc**	**330,136**	**15,496**	**14,423**	**93**	**22.9**	**58.1**
Arabidopsis	**TAIR**	**171,409**	**33,200**	**30,193**	**91**	**5.7**	**35.5**
Caenorhabditis	**WormBase**	**112,725**	**28,202**	**23,861**	**85**	**4.7**	**34.2**
Drosophila	**FlyBase**	**185,863**	**21,736**	**20,259**	**93**	**9.2**	**40.9**
Danio rerio	**Ensembl**	**214,809**	**27,305**	**25,918**	**95**	**8.3**	**35.9**
Homo sapiens	**Ensembl**	**670,597**	**71,913**	**66,170**	**92**	**10.1**	**37.1**

[1] Data Source URLs: SGD - http://www.yeastgenome.org/

ChlamyCyc - http://chlamyto.mpimp-golm.mpg.de/chlamycyc/index.jsp

TAIR - http://www.arabidopsis.org/

Wormbase - http://www.wormbase.org/

FlyBase - http://flybase.org/

Ensembl *Danio rerio* - http://uswest.ensembl.org/Danio_rerio/Info/Index

Ensembl *Homo sapiens*- http://uswest.ensembl.org/Homo_sapiens/Info/Index

[2] "Transcripts" refers to protein encoding transcripts mapped onto chromosomes (i.e., scaffolds are not included).

Note: For *Chlamydomonas reinhardtii* the average number of ZFN target sites per transcript is very high. This likely reflects the increased GC content of this genome.

Resources available in ZFNGenome

Users can choose the model organism of interest from the ZFNGenome homepage, http://bindr.gdcb.iastate.edu/ZFNGenome, by choosing an organism from the left hand column of the front page or via the "Data Source" dropdown menu from within an organism's ZFNGenome page (Figure. 3.3 A and 3.3 B). Figure 3.3 B is a screenshot of the output displayed in response to a search for ZFN target sites in the *Saccharomyces cerevisiae* genome. Several standard GBrowse tracks are displayed by default (genes, transcripts, coding regions, etc.). The OPEN Zinc Finger Nuclease Sites track shows that within the 2.187 kb region illustrated (gene YLR219W) there are 17 potential ZFN sites located within the coding region of this gene. Because OPEN reagents are available to recognize all possible GNN and some TNN triplets, we have included a track illustrating analysis of the GC content of the DNA. Clicking on any genomic feature illustrated below the sequence reveals additional information about that feature. For example, clicking on the "OPEN Zinc Finger Nuclease Site" (AGCAGCGTCNNNNNNNGAAGGTGTG) opens a page containing more information about the site, as illustrated in Figure 3.3 D. The "Note" sections on this page provide links to ZiFDB [40], a repository for zinc finger arrays that have been experimentally validated, and ZiFiT [37], a tool for identifying potential ZFP and ZFN target sites. A hyperlink to NCBI BLAST [41] can be used to check whether the identified ZFN site is unique in the genome of interest. By clicking on other features, an investigator can access the exact sequence, chromosomal position and sources of reagents needed to experimentally target the chosen site. Users may customize the GBrowse display by choosing which feature(s) to display (using the -/+ buttons on the left), and defining the order in which features are displayed by dragging and dropping the features within the browser window. The ZFN tracks can be exported back into the "home" GBrowse website for a model organism by clicking on the "share the track" button (details provided in the Tutorial, Figure 3.3 C). Users can also utilize Help, Instruction, and Tutorial functions within the browser windows to obtain more information about navigating ZFNGenome.

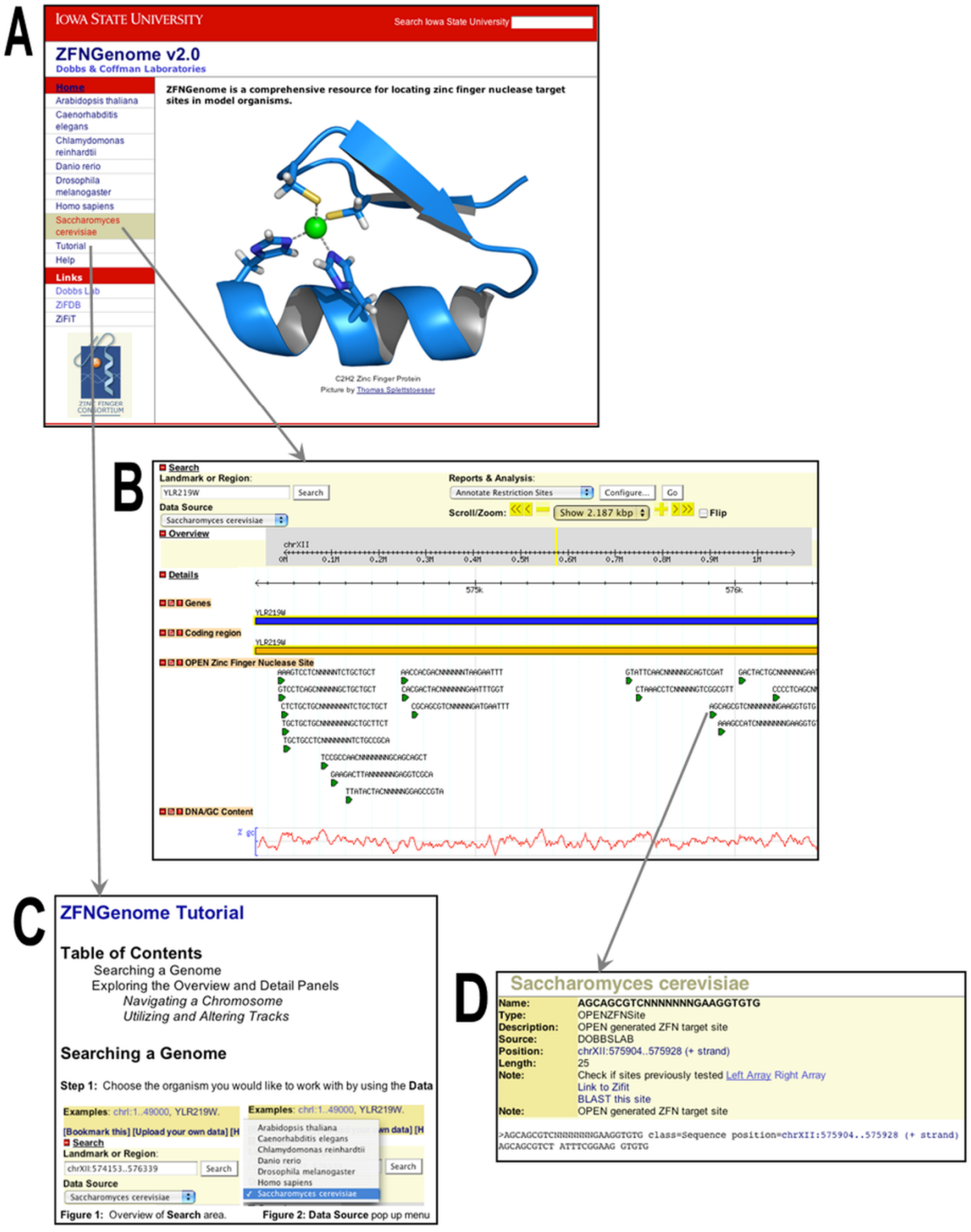


Figure 3.3. Examples of resources available in ZFNGenome. (A) The ZFNGenome Homepage is shown. From here, the user can select a model organism from the seven shown in the left hand

column. In addition, links to the ZFNGenome *Tutorial* and *Help* pages are provided. (B) A screenshot of the result of a search of the *S. cerevisiae* gene YLR219W is displayed. Key areas of the browser include the search box and the "Scroll/Zoom" areas at the top. The "Overview" and "Detail" panels serve as controls for visualizing the genome. This search shows the single coding region of this gene has 17 potential ZFN target sites, color-coding according to their "uniqueness" and "ZiFOpT" scores (see text). Additional information on each of the tracks can be obtained by clicking on details of the track. For example, clicking on one of the OPEN Zinc Finger Nuclease Sites links the user to details about that specific ZFN target. (C) The ZFNGenome *Tutorial* offers instructions on navigating the database. The *Tutorial* can be accessed from the Homepage or from any GBrowse page within ZFNGenome. *Help* and *Instruction* links are provided from the GBrowse pages. (D) Clicking on a ZFN target site opens a new window that provides links to ZiFDB, which provides additional information for each zinc finger array, ZiFiT the zinc finger design software that includes the OPEN design method and zinc finger pools, and the BLAST server at National Center for Biotechnology Information (NCBI).

To evaluate the reliability of data presented in ZFNGenome, we compared our results with other published data. Two types of data are presented in ZFNGenome: annotated genomic features and potential ZFN target sites. The sources from which we acquired the genomic features are listed in Table 1**.** These are widely considered to be the “gold standard” data sources for the model organisms analyzed because they are carefully annotated and repeatedly evaluated by the curators and users of these databases. These source databases are also extensively used by investigators utilizing the various model organisms and are therefore familiar to users. To identify potential errors that may have been introduced during pre-processing or data analysis, we performed quality assurance tests as follows: 1) for each organism, several 5 kb segments of genomic sequence were randomly selected from each chromosome; 2) selected chromosomal DNA sequences were individually re-scanned using the ZiFiT web server [37] to identify potential OPEN ZFN sites; 3) sites identified by the ZiFiT server were directly compared to the results for the corresponding region obtained from the ZFNGenome database; genomic features were checked against the original database. To improve the user interface and documentation, we incorporated suggestions from at least one expert scientist for each of model organisms included in ZFNGenome.

DISCUSSION

Currently available ZFNs can target 85 - 95% of protein coding transcripts in 7 model organisms.

The results presented in Table 1 illustrate both the power and current limitations of OPEN ZFN engineering technology and identify gaps where further improvement is needed. Most striking is the relatively high level of coverage currently achievable. This ranges from 85% of protein coding transcripts in *Caenorhabditis elegans* to 95% of protein coding transcripts in *Danio rerio.* Also noteworthy is the number of potential target sites available within any given transcript: in the model organisms examined to date, each transcript contains, on average 5 – 23 target sites (Table 1). The current lack of OPEN ZFN reagents for targeting TNN, ANN and CNN triplets is a limitation, especially in organisms with AT rich genomes. However, even in Arabidopsis (35.5% GC) 91% of the protein coding transcripts are potentially targetable. As more ZFN reagents for targeting additional triplets become available, the applicability of ZFN technology will continue to increase.

The first study in which the entire genome of a model organism was analyzed to identify potential target sites for ZFNs focused on the zebrafish, *Danio rerio* [6]. In that study, identified ZFN target sites were published in the form of 26 supplemental tables (one for each chromosome). Although this information has apparently proven useful for members of the zebrafish community, ZFNGenome was developed in an effort to make such large datasets searchable and more readily accessible to a broader group of researchers working in zebrafish as well as other model organisms.

In the first implementation of ZFNGenome, we used GBrowse version 1.67 with a BerkeleyDB back end to display all potential ZFN target sites found in *Arabidopsis* [15]. A total of 381,497 sites were identified, 171,409 of which were located within coding regions (an average of 5.7 sites per protein coding transcript). The current version of ZFNGenome (2.0) has been expanded to include *S. cerevisiae*, *C. reinhardtii*, *C. elegans*, *D. melanogaster*, *D. rerio*, and *H. sapiens*. In addition, it has been implemented in the newer GBrowse 1.7 with a MySQL database, which results in a more dynamic and user-friendly interface. GBrowse 1.7 is a robust and highly customizable browser available from the Generic Model Organism Database project (GMOD) [36]. A noteworthy feature is the ability to share tracks with other GBrowse-based resources. To date ~119 implementations of GBrowse are available (http://gmod.org/wiki/GMOD_Users). Users accustomed to using popular model organism resources, such as TAIR for Arabidopsis [42] or FlyBase for Drosophila [43], can simply

export tracks containing ZFN target sites from ZFNGenome and into their browser of choice for further analysis.

Related Resources

Several existing databases house information on ZFPs and associated binding sites. ZiFDB (http://bindr.gdcb.iastate.edu/ZiFDB) contains information about engineered zinc finger arrays and individual modules that have been experimentally evaluated for function *in vivo* [40]. ZifBase (http://web.iitd.ac.in/~sundar/zifbase/) is a repository that includes information about both naturally occurring and engineered zinc finger proteins [44]. Sequences of ZFP binding sites are also collected in TRANSFAC [45] (http://www.gene-regulation.com/pub/databases.html) and JASPAR (http://jaspar.genereg.net/) [46]. Tools for predicting the DNA target sites for a selected ZFP include ZIFIBI (http://bioinfo.hanyang.ac.kr/ZIFIBI/frameset.php), a hidden Markov model based predictor that takes into account the interdependence between positions -1, +3 and +6 of a chosen ZFP to predict its potential DNA binding site(s) [47]. Also, Persikov *et al*. [48] have used support vector machines (SVMs) to predict and rank potential ZFP binding sites for a selected ZFP.

Several web-based tools for identifying potential ZFN binding sites within a given DNA sequence are currently available. Zinc Finger Tools (http://www.scripps.edu/mb/barbas/zfdesign/zfdesignhome.php) can be used to identify target sites for zinc finger arrays composed of available modules (16 GNN, 15 ANN, 15 CNN) generated by the Barbas laboratory, within any given DNA sequence up to 10 kb in length [49]. ZifBase tools (http://web.iitd.ac.in/~sundar/zifbase/) can identify target sites in a given DNA sequence, with the option of using target site triplet composition (i.e., the number of GNN, CNN, TNN and ANN triplets), as a selection criterion. TagScan (http://www.isrec.isb-sib.ch/tagger/tagscan.html) is capable of performing searches for either exact or nearly exact matches ($\leq$ 2 mismatches) between a given query sequence, such as a ZFP target site, and a large database, such as a genomic sequence database [50]. ZiFiT (http://bindr.gdcb.iastate.edu/zifit/) is similar to ZFTools in that it allows users to identifying target sites for ZFNs. ZiFiT also can identify sites potentially targetable with ZFPs made from zinc finger modules developed and/or characterized by the Barbas lab, Sangamo

BioSciences, Inc., and Toolgen (http://www.toolgen.com). In contrast to all of these existing web-based tools, which identify potential ZFN target sites within a user-provided DNA sequence (typically < 10 kb), ZFNGenome is a comprehensive repository that contains all potential ZFN sites targetable using available OPEN reagents in the complete genomic sequences of 7 model organisms.

Planned future development

ZFNGenome will be updated regularly to incorporate revisions in genomic DNA sequences and annotations, and to take into account new potential ZFN target sites that can be considered when new reagents, such as additional OPEN pools, become available. The genomes of several other established and emerging model organisms currently in the pipeline include: maize, rice, soybean, red flower beetle, mouse, and rat. We also intend to implement additional features, including capabilities for identifying target sites for ZFNs made by other publicly available engineering methods (e.g., modular assembly). Finally, because the experimental generation and testing of ZFNs using the OPEN protocol is not a trivial undertaking, the utility of a method to discriminate between ZFN target sites that are likely to function successfully *in vivo* and those that are not, cannot be over-emphasized. Our analysis indicates that, on average, every transcript in the zebrafish genome contains ~ 8 potential ZFN target sites (see Table 1). Thus, software for reliably ranking identified potential target sites according to the probability that they will function successfully *in vivo* is important for improving the time and cost-effectiveness of genomic modification experiments utilizing ZFNs. The next version of ZFNGenome will include an experimentally validated scoring scheme, ZiFOpT (J. Sander and D. Reyon, personal communication) to provide users with a ranked list of specific DNA sites that are most amenable to ZFN targeting.

CONCLUSIONS

OPEN is a robust, publicly available, experimental platform for the generation of engineered ZFNs that function with high specificity *in vivo*. ZFNGenome was developed to enhance and broaden the applicability of ZFNs for genomic modification by providing an online resource that contains all potential target sites for OPEN-generated ZFNs in the sequenced genomes of several model organisms. ZFNGenome has a user-friendly interface

and is seamlessly integrated with other publicly available Zinc Finger Consortium resources, such as ZiFiT and ZiFDB. ZFNGenome should be a valuable resource for scientists and clinicians who wish to exploit the powerful technologies for genome modification now available as a result of recent developments in ZFP design and engineering.

AVAILABILITY AND REQUIREMENTS

ZFNGenome is freely available over the web at *http://bindr.gdcb.iastate.edu/ZFNGenome.*

LIST OF ABBREVIATIONS USED

OPEN = Oligomerized Pool ENgineering

ZF = Zinc Finger

ZFA = Zinc Finger Array

ZFP = Zinc Finger Protein

ZFN = Zinc Finger Nuclease

AUTHORS' CONTRIBUTIONS

All authors contributed to the overall concept and DR directed the design and implementation of the database. DR and JK identified ZFN target sites in model organism genomes, constructed the database, provided online documentation, and designed the web interface. JK, DR, and CC performed quality control assessment. DR, JK, DD and CC drafted the manuscript. All authors read, contributed to revisions of and approved the final manuscript.

ACKNOWLEDGEMENTS

We thank members of our research groups for helpful discussions and Chris Campbell for assistance with 64 bit conversion. We also thank Jo Anne Powell-Coffman, Jeff Essner, David Wright, Ben Lewis and Rasna Walia, for critical comments on the ZFNGenome server and the manuscript. This work was supported by NSF DBI 0923827 to DFV, DD, and JKJ, NIH grants R01 GM069906 and R01 GM088040 to JKJ, The Roy J. Carver Charitable Trust

08-3185 to CRC, and the Center for Integrated Animal Genomics at Iowa State University to DD. JDS was supported by the NIH T32 CA009216.

REFERENCES

1. Carroll, D., *Progress and prospects: zinc-finger nucleases as gene therapy agents.* Gene Ther, 2008. **15**(22): p. 1463-8.
2. Cathomen, T. and J.K. Joung, *Zinc-finger nucleases: the next generation emerges.* Mol Ther, 2008. **16**(7): p. 1200-7.
3. Urnov, F.D., et al., *Genome editing with engineered zinc finger nucleases.* Nat Rev Genet, 2010. **11**(9): p. 636-46.
4. Beumer, K., et al., *Efficient gene targeting in Drosophila with zinc-finger nucleases.* Genetics, 2006. **172**(4): p. 2391-403.
5. Doyon, Y., et al., *Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(6): p. 702-8.
6. Foley, J.E., et al., *Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN).* PLoS ONE, 2009. **4**(2): p. e4348.
7. Geurts, A.M., et al., *Knockout rats via embryo microinjection of zinc-finger nucleases.* Science, 2009. **325**(5939): p. 433.
8. Lee, H.J., E. Kim, and J.S. Kim, *Targeted chromosomal deletions in human cells using zinc finger nucleases.* Genome Res, 2010. **20**(1): p. 81-9.
9. Maeder, M.L., et al., *Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification.* Mol Cell, 2008. **31**(2): p. 294-301.
10. Osakabe, K., Y. Osakabe, and S. Toki, *Site-directed mutagenesis in Arabidopsis using custom-designed zinc finger nucleases.* Proc Natl Acad Sci U S A, 2010. **107**(26): p. 12034-9.
11. Perez, E.E., et al., *Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(7): p. 808-16.
12. Shukla, V.K., et al., *Precise genome modification in the crop species Zea mays using zinc-finger nucleases.* Nature, 2009. **459**(7245): p. 437-41.
13. Townsend, J.A., et al., *High-frequency modification of plant genes using engineered zinc-finger nucleases.* Nature, 2009. **459**(7245): p. 442-5.
14. Voigt, B. and T. Serikawa, *Pluripotent stem cells and other technologies will eventually open the door for straightforward gene targeting in the rat.* Dis Model Mech, 2009. **2**(7-8): p. 341-3.
15. Zhang, F., et al., *High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases.* Proc Natl Acad Sci U S A, 2010. **107**(26): p. 12028-33.
16. Zou, J., et al., *Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells.* Cell Stem Cell, 2009. **5**(1): p. 97-110.
17. Kaiser, J., *Gene therapy. Putting the fingers on gene repair.* Science, 2005. **310**(5756): p. 1894-6.

18. Pearson, H., *Protein engineering: The fate of fingers.* Nature, 2008. **455**(7210): p. 160-4.
19. Scott, C.T., *The zinc finger nuclease monopoly.* Nat Biotechnol, 2005. **23**(8): p. 915-8.
20. Kim, Y.G., J. Cha, and S. Chandrasegaran, *Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain.* Proc Natl Acad Sci U S A, 1996. **93**(3): p. 1156-60.
21. Kim, Y.G., et al., *Site-specific cleavage of DNA-RNA hybrids by zinc finger/FokI cleavage domain fusions.* Gene, 1997. **203**(1): p. 43-9.
22. Klug, A., *Towards therapeutic applications of engineered zinc finger proteins.* FEBS Lett, 2005. **579**(4): p. 892-4.
23. Mani, M., et al., *Binding of two zinc finger nuclease monomers to two specific sites is required for effective double-strand DNA cleavage.* Biochem Biophys Res Commun, 2005. **334**(4): p. 1191-7.
24. Bibikova, M., et al., *Enhancing gene targeting with designed zinc finger nucleases.* Science, 2003. **300**(5620): p. 764.
25. Hockemeyer, D., et al., *Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases.* Nat Biotechnol, 2009. **27**(9): p. 851-7.
26. Lombardo, A., et al., *Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery.* Nat Biotechnol, 2007. **25**(11): p. 1298-306.
27. Urnov, F.D., et al., *Highly efficient endogenous human gene correction using designed zinc-finger nucleases.* Nature, 2005. **435**(7042): p. 646-51.
28. Bibikova, M., et al., *Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases.* Genetics, 2002. **161**(3): p. 1169-75.
29. Meng, X., et al., *Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(6): p. 695-701.
30. Hurt, J.A., et al., *Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12271-6.
31. Maeder, M.L., et al., *Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays.* Nat Protoc, 2009. **4**(10): p. 1471-501.
32. Joung, J.K., D.F. Voytas, and T. Cathomen, *Reply to "Genome editing with modularly assembled zinc-finger nucleases".* Nat Methods, 2010. **7**(2): p. 91-2.
33. Kim, H.J., et al., *Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly.* Genome Res, 2009. **19**(7): p. 1279-88.
34. Ramirez, C.L., et al., *Unexpected failure rates for modular assembly of engineered zinc fingers.* Nat Methods, 2008. **5**(5): p. 374-5.
35. Foley, J.E., et al., *Targeted mutagenesis in zebrafish using customized zinc-finger nucleases.* Nat Protoc, 2009. **4**(12): p. 1855-67.
36. Stein, L.D., et al., *The generic genome browser: a building block for a model organism system database.* Genome Res, 2002. **12**(10): p. 1599-610.
37. Sander, J.D., et al., *ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool.* Nucleic Acids Res, 2010. **38 Suppl**: p. W462-8.

38. Sander, J.D., et al., *Zinc Finger Targeter (ZiFiT): an engineered zinc finger/target site design tool.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W599-605.
39. Stajich, J.E., et al., *The Bioperl toolkit: Perl modules for the life sciences.* Genome Res, 2002. **12**(10): p. 1611-8.
40. Fu, F., et al., *Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays.* Nucleic Acids Res, 2009. **37**(Database issue): p. D279-83.
41. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
42. Swarbreck, D., et al., *The Arabidopsis Information Resource (TAIR): gene structure and function annotation.* Nucleic Acids Res, 2008. **36**(Database issue): p. D1009-14.
43. Tweedie, S., et al., *FlyBase: enhancing Drosophila Gene Ontology annotations.* Nucleic Acids Res, 2009. **37**(Database issue): p. D555-9.
44. Jayakanthan, M., et al., *ZifBASE: a database of zinc finger proteins and associated resources.* BMC Genomics, 2009. **10**: p. 421.
45. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
46. Bryne, J.C., et al., *JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D102-6.
47. Cho, S.Y., et al., *ZIFIBI: Prediction of DNA binding sites for zinc finger proteins.* Biochem Biophys Res Commun, 2008. **369**(3): p. 845-8.
48. Persikov, A.V., R. Osada, and M. Singh, *Predicting DNA recognition by Cys2His2 zinc finger proteins.* Bioinformatics, 2009. **25**(1): p. 22-9.
49. Mandell, J.G. and C.F. Barbas, 3rd, *Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W516-23.
50. Iseli, C., et al., *Indexing strategies for rapid searches of short words in genome sequences.* PLoS ONE, 2007. **2**(6): p. e579.

CHAPTER 4. TARGETED GENE DISRUPTION IN SOMATIC ZEBRAFISH CELLS USING ENGINEERED TALENS

Jeffry D Sander, Lindsay Cade, Cyd Khayter, Deepak Reyon, Randall T Peterson, J Keith Joung & Jing-Ruey J Yeh. *Nature Biotechnology.* 2011 Aug 5.

To the Editor:

Miller et al. recently described a transcription activator–like effector nuclease (TALEN) architecture for efficient genome editing in cultured human cells [1]. We sought to determine whether the same framework could be used to efficiently disrupt endogenous genes in somatic cells of zebrafish and how the efficiency of TALENs compares with that obtained using engineered zinc-finger nucleases (ZFNs).

TALENs, which comprise an engineered array of transcription activator–like effector repeats fused to the nonspecific FokI cleavage domain, introduce targeted double-stranded breaks in human cells with high efficiency. Repair of these double-stranded breaks by normal DNA repair mechanisms, such as Non-Homologous End-Joining (NHEJ) or homologous recombination, enables introduction of alterations at or near the site of the break. A single 34-amino-acid transcription activator–like effector repeat binds to one bp of DNA, and repeats which bind each of the four DNA bases have been described [2, 3]. These modules can be assembled into arrays capable of binding extended DNA sequences. TALENs may have advantages over engineered ZFNs due to the relative ease with which they can be designed and their potential ability to be targeted to a wide range of sequences, with target sites reported to be as frequent as 1 in 35 bp of random DNA sequence [4].

Previous studies have shown that error-prone repair of ZFN-induced double-stranded breaks by NHEJ can result in the efficient introduction of small insertions or deletions (indels) at cleavage sites in endogenous zebrafish genes [5-7]. These indels frequently result in frameshift knockout mutations that can be passed through the germ line to create mutant fish [5-9]. ZFN technology has enabled reverse genetics studies to be performed in zebrafish. However, engineering ZFNs can be challenging because of the need to account for context-dependent effects among individual fingers in an array. In addition, although many zebrafish genes can be targeted with ZFNs made by publicly available methods that account for

context dependence [6, 10], it can be difficult to target sequences within some genes in zebrafish due to the currently limited targeting range of publicly available ZFN engineering platforms. Thus, use of TALENs for targeted mutation in zebrafish could provide an important additional capability for this model organism.

To test the ability of TALENs to function in zebrafish, we targeted DNA sequences in two endogenous zebrafish genes, gria3a and hey2 (Fig. 4.1). To avoid confounding effects that might affect binding and cleavage of DNA sites by TALENs (e.g., chromatin structure or DNA methylation), we chose to target sequences that we had efficiently altered previously in zebrafish using engineered ZFNs (Supplementary Figs. 1 and 2). Using an iterative assembly approach (Supplementary Methods, Supplementary Figs. 3–10, and Supplementary Table 1), we constructed four TALEN monomers designed to target partially overlapping sites in gria3a and two TALEN monomers designed to target a site in hey2 (Fig. 4.1 and Supplementary Fig. 3). These six TALEN monomers all harbor the wild-type FokI cleavage domain (Supplementary Figs. 4 and 5) and can be paired in combinations to make three TALEN dimers to the gria3a gene and one TALEN dimer to the hey2 gene (Fig. 4.1). We injected RNAs encoding the various TALEN pairs into one-cell-stage zebrafish embryos and determined the frequency of NHEJ-mediated mutagenesis at the target site by sequence analysis of alleles from pooled injected embryos (Supplementary Methods). We found that all four pairs of TALENs induced targeted indels with high mutation frequencies ranging from 11% to 33% (Fig. 4.1). These frequencies are comparable to what we obtained using ZFNs targeted to DNA sequences in the same vicinity of the gene (Supplementary Fig. 1); however, we note that the TALENs harbor wild-type FokI domains whereas the ZFNs harbor obligate heterodimeric FokI domains [11]. Although small indels were typically observed with the TALENs, we also observed deletions as large as 303 bp (Fig. 4.1).

hey2:

#1297/ #1257 Mutations in 12 of 110 sequences: ~11%

```
GCTCTTCCGTTTCCACATCCACCACATCCCAACAGAGCAGCGGGAGCAGCAGTAAACC  WT

<------------------------------------------GAGCAGCAGTAAACC  Δ142
GCTCTTCCGTTTCCACATCCACC--------------CAGCGGGAGCAGCAGTAAACC  Δ14
GCTCTTCCGTTTCCACATCCACC--------------ACAGCGGGAGCAACAGTAAACC  Δ13
GCTCTTCCGTTTCCACATCCAC-----------AGAGCAGCGGGAGCAGCAGTAAACC  Δ11              [2x]
GCTCTTCCGTTTCCACATCCACCAC--------AGAGCAGCGGGAGCAGCAGTAAACC  Δ8               [3x]
GCTCTTCCGTTTCCACATCCACCACAT-------tGAGCAGCGGGAGCAACAGTAAACC  Δ6 (Δ7 and +1)
GCTCTTCCGTTTCCACATCCACCACATC--AACAGAGCAGCGGGAGCAGCAGTAAACC  Δ2
GCTCTTCCGTTTCCACATCCACCACATaaaccaccacACAGAGCAGCGGGAGCAGCAG  +6 (Δ4 and +10)

ACCTTCCCTCTATCATT<----------/ /---------->TCTGGGAAGAAAAGAAA  Δ303
```

gria3a:

#1258/ #1260 Mutations in 13 of 89 sequences: ~15%

```
GGAGTCGTCCAATAGCTTCTCAGTCACGCACGCCTGTGAGTTTCTGCTCTTTATCTT  WT

GGAGTCGTCCAATAGCTTC------------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ12
GGAGTCGTCCAATAGCTTCTCA---------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ9
GGAGTCGTCCAATAGCTTCTCAGT---------CTGTGAGTTTCTGCTCTTTATCTT  Δ9
GGAGTCGTCCAATAGCTTCTCAG---------CCTGTGAGTTTCTGCTCTTTATCTT  Δ9               [2x]
GGAGTCGTCCAATAGCTTCTCAGTCA-----GCCTGTGAGTTTCTGCTCTTTATCTT  Δ5
GGAGTCGTCCAATAGCTTCTCAGTCAgaaa--CCTGTGAGTTTCTGCTCTTTATCTT  Δ2 (Δ6 and +4)
GGAGTCGTCCAATAGCTTCTCAGTCtcagtCGCCTGTGAGTTTCTGCTCTTTATCTT  +0 (Δ5 and +5)
GGAGTCGTCCAATAGCTTCTCAGTCacgcACGCACGCCTGTGAGTTTCTGCTCTTTA  +4               [3x]
GGAGTCGTCCAATAGCTTCTCAGctgtgcctgtaACGCCTGTGAGTTTCTGCTCTTT  +5 (Δ6 and +11) [2x]
```

#1258/ #1259 Mutations in 21 of 68 sequences: ~31%

```
GGAGTCGTCCAATAGCTTCTCAGTCACGCACGCCTGTGAGTTTCTGCTCTTTATCTT  WT

GGAGTCGTCCAATAGCT------------------GTGAGTTTCTGCTCTTTATCTT  Δ18
GGAGTCGTCCAATAGCTTCTC----------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ10              [3x]
GGAGTCGTCCAATAGCTTCTCAGT---------CTGTGAGTTTCTGCTCTTTATCTT  Δ9               [3x]
GGAGTCGTCCAATAGCTTCTCAGTCA---------GTGAGTTTCTGCTCTTTATCTT  Δ9
GGAGTCATCCAATAGCTTCTCAGTC------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ6               [5x]
GGAGTCATCCAATAGCTTCTCAGTCAgaa------GTGAGTTTCTGCTCTTTATCTT  Δ6 (Δ9 and +3)
GGAGTCGTCCAATAGCTTCTCAGTCt-----CGCCTGTGAGTTTCTGCTCTTTATCTT  Δ4 (Δ5 and +1)  [3x]
GGAGTCGTCCAATAGCTTCTCAGTCA----CGCCTGTGAGTTTCTGCTCTTTATCTT  Δ4
GGAGTCGTCCAATAGCTTCTCAGcttct--CGCCTGTGAGTTTCTGCTCTTTATCTT  Δ2 (Δ7 and +5)
GGAGTCGTCCAATAGCTTCTCAGTCAgt--CGCCTGTGAGNNTCTGNTCTTTATCTT  Δ2 (Δ4 and +2)
GGAGTCGTCCAATAGCTTCTCAGTttctcagcttCGCCTGTGAGTTTCTGCTCTTTA  +4 (Δ6 and +10)
```

#1295/ #1260 Mutations in 26 of 79 sequences: ~33%

```
GGAGTCGTCCAATAGCTTCTCAGTCACGCACGCCTGTGAGTTTCTGCTCTTTATCTT  WT

GG-----------------------------------AGTTTCTGCTCTTTATCTT  Δ36
GGAGTCGTCCAATAGCTTC------------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ12
GGAGTCGTCCAATAGCTTCTC----------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ10              [2x]
GGAGTCGTCCAATAGCTTCTCAGT---------CTGTGAGTTTCTGCTCTTTATCTT  Δ9               [3x]
GGAGTCGTCCAATAGCTTCTt----------CGCCTGTGAGTTTCTGCTCTTTATCTT  Δ9 (Δ10 and +1)
GGAGTCGTCCAATAGCTTCTCA---------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ9
GGAGTCGTCCAATAGCTTCTCAGTCA---------GTGAGTTTCTGCTCTTTATCTT  Δ9               [2x]
GGAGTCGTCCAATAGCTTCTCAGTC--------CTGTGAGTTTCTGCTCTTTATCTT  Δ8               [2x]
GGAGTCGTCCAATAGCTTCTCAGTC-------CCTGTGAGTTTCTGCTCTTTATCTT  Δ7
GGAGTCGTCCAATAGCTTCTCAGTCA-------CTGTGAGTTTCTGCTCTTTATCTT  Δ7               [3x]
GGAGTCGTCCAATAGCTTCTCAGTCA------CCTGTGAGTTTCTGCTCTTTATCTT  Δ6
GGAGTCGTCCAATAGCTTCTCAGTCACG-----TGTGAGTTTCTGCTCTTTATCTT  Δ6
GGAGTCATCCAATAGCTTCACAGTC------GCCTGTGAGTTTCTGCTCTTTATCTT  Δ6               [2x]
GGAGTCATCCAATAGCTTCTCAGTCACGC----CTGTGAGTTTCTGCTCTTTATCTT  Δ4               [2x]
GGAGTCGTCCAATAGCTTCTCAGTCAga---GCCTGTGAGTTTCTGCTCTTTATCTT  Δ3 (Δ5 and +2)
GGAGTCATCCAATAGCTTCTCAGTCACGCACGcacgCCTGTGAGTTTCTGCTCTTTA  +4
GGAGTCGTCCAATAGCTTCTCAGTCAgtgagtcacactCGCCTGTGAGTTTCTGCTC  +8 (Δ4 and +12)
```

Figure 4.1: Target sequences, frequencies of mutations and mutations induced by TALENs in embryonic zebrafish cells. For each pair of TALENs, the wild-type (WT) target sequence is shown at the top with the intended target sites of the TALENs marked in yellow. Deletions are indicated by red dashes against a gray background, and insertions by lowercase blue letters against a light blue background. The sizes of the insertions (+) or deletions (Δ) are indicated to the right of each mutant allele. The number of times that each mutant allele was isolated is shown between square brackets. Mutation frequencies are calculated as the number of mutant alleles isolated/the total number of alleles analyzed. For the *hey2* gene, we also identified two larger deletions 142 and 303 bp in length, which extend substantially beyond the intended target sites of the TALENs.

To assess the toxicity of our engineered TALENs, we scored the percentages of dead and deformed embryos that resulted from mRNA microinjections (Supplementary Fig. 11). Although we cannot directly compare these results with the microinjections of ZFNs due to the differences between the FokI endonuclease domains used (EL/KK heterodimeric FokI11 for ZFNs versus wild-type FokI for the TALENs) and the specific sequences targeted, the toxicity we observed with injections of 600 pg of TALEN mRNAs (range of 40–80%) appears similar to that observed with 400–500 pg of mRNAs encoding ZFNs targeted to sequences in the same vicinity (Supplementary Fig. 12) and to other genes [6].

Successful germline transmission of these mutations will be critical for using TALENs to perform reverse genetics in zebrafish, although further experiments are needed both to demonstrate this and to evaluate the frequency and range of TALEN-induced off-target effects. Given that the frequencies of mutation and the extent of toxicities we observe are similar to what we have seen with ZFNs, we expect that TALEN-induced mutations should be efficiently passed through the germ line. Progeny bearing TALEN-induced mutations, which unlike founder F0 fish would be uniformly mutated in all cells, will reveal whether both mono-allelic and bi-allelic alterations of a gene are possible and provide a more straightforward background for analysis of off-target effects.

In summary, we have shown that the TALEN framework described by Miller et al.[1] can be used to efficiently introduce targeted indel mutations in endogenous genes of zebrafish somatic cells. Although in this study we chose two genomic loci that have been successfully targeted with ZFNs before, all six TALEN monomers we constructed showed high mutagenesis activities when tested in various pairwise combinations. This suggests that the TALEN framework is also highly robust and effective in zebrafish. As is the case with

ZFNs, the complete genome-wide spectrum of off-target mutations introduced by TALENs remains unknown. However, expression of the TALENs we made in zebrafish did not show toxicity substantially different from that observed with expression of ZFNs, suggesting that the magnitude of off-target effects may be comparable between the two classes of nuclease. In principle, off-target mutations generated by TALENs can be removed by outcrossing the founder, provided that the alterations are not tightly linked to the intended mutation. Moreover, mutant phenotypes could also be confirmed by generation of a second mutant allele using nucleases targeted to a different site. TALENs may offer potential advantages over ZFNs for mutagenesis of genes in zebrafish and other model organisms such as Caenorhabditis elegans [12], because they can be easily and quickly assembled in a modular fashion and can potentially target a greater range of DNA sequences. Thus, we expect that the ability to use both ZFNs and TALENs should enable any researcher to rapidly and easily create targeted mutations in any zebrafish gene of interest.

ACKNOWLEDGEMENTS

We thank D. Voytas and A. Bogdanove for helpful discussions in the early stages of this project and J.Foley, M. Maeder, S. Thibodeau-Beganny, F. Zhang, M. Christian and D. Voytas for help with characterizing the *hey2*- and *gria3a*-targeted ZFN pairs. This work was supported by US National Institutes of Health (NIH) R01 GM088040 (J.K.J. & R.T.P.), NIH Director's Pioneer Award DP1OD006862 (J.K.J.), NIH T32 CA009216 (J.D.S.), NIH K01 AG031300 (J.-R.J.Y.), the Jim and Ann Orr Massachusetts General Hospital (MGH) Research Scholar award (J.K.J.), and the Charles and Ann Sanders MGH Research Scholar award (R.T.P.).

CONFLICT OF INTEREST STATEMENT:

None declared.

REFERENCES

1. Miller, J.C., et al., *A TALE nuclease architecture for efficient genome editing.* Nat Biotechnol, 2011. **29**(2): p. 143-8.
2. Boch, J., et al., *Breaking the code of DNA binding specificity of TAL-type III effectors.* Science, 2009. **326**(5959): p. 1509-12.

3. Moscou, M.J. and A.J. Bogdanove, *A simple cipher governs DNA recognition by TAL effectors.* Science, 2009. **326**(5959): p. 1501.
4. Cermak, T., et al., *Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting.* Nucleic Acids Res.
5. Doyon, Y., et al., *Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(6): p. 702-8.
6. Foley, J.E., et al., *Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN).* PLoS ONE, 2009. **4**(2): p. e4348.
7. Meng, X., et al., *Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(6): p. 695-701.
8. Cifuentes, D., et al., *A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity.* Science. **328**(5986): p. 1694-8.
9. Siekmann, A.F., et al., *Chemokine signaling guides regional patterning of the first embryonic artery.* Genes Dev, 2009. **23**(19): p. 2272-7.
10. Sander, J.D., et al., *Selection-free zinc-finger-nuclease engineering by context-dependent assembly (CoDA).* Nat Methods, 2011. **8**(1): p. 67-9.
11. Miller, J.C., et al., *An improved zinc-finger nuclease architecture for highly specific genome editing.* Nat Biotechnol, 2007. **25**(7): p. 778-85.
12. Wood, A.J., 2011. p. 307.

CHAPTER 5. USER-FRIENDLY PROTOCOL AND SOFTWARE FOR RAPID ENGINEERING OF DESIGNER TALE NUCLEASES (TALENS)

Jeffry D. Sander*, Deepak Reyon*, Cyd Khayter, & J. Keith Joung. *Nature Protocols* (Under Review)

* Co-first authors

ABSTRACT

Engineered transcription activator-like effector nucleases (TALENs) are broadly useful tools for performing targeted genome editing in a wide variety of organisms and cell types including zebrafish, C. elegans, rat, human somatic cells, and human pluripotent stem cells. We recently described a serial, hierarchical ligation strategy for rapidly assembling TALENs that requires neither PCR nor specialized multi-fragment ligations and can therefore be implemented by any laboratory. Here we provide a detailed protocol for practicing our assembly method together with user-friendly, web-based software that both identifies potential target sites in sequences of interest and generates printable graphical guides that facilitate assembly of TALENs. All plasmids required to perform our assembly method are publicly available through the Addgene plasmid distribution service (http://www.addgene.org/talengineering). With the platform of reagents, protocols, and software we describe, researchers can easily engineer multiple TALENs in two weeks or less using standard cloning techniques.

INTRODUCTION

Engineered transcription activator-like effector nucleases (TALENs) have recently generated much interest as a broadly applicable technology for highly efficient genome editing [1]. TALENs, like zinc finger nucleases (ZFNs), are customizable restriction enzymes consisting of an engineered DNA-binding domain fused to a non-specific nuclease domain [2-6]. Site-specific double-stranded DNA breaks induced by TALENs have been used to introduce sequence alterations at investigator-specified endogenous genes in a variety of

organisms and cell types (including yeast,[7] plants,[8] *C. elegans*,[9] zebrafish,[10, 11] rats,[12] and human somatic[3, 6, 8] and pluripotent stem cells[13]).

The DNA-binding domain of a TALEN consists of an array of repeat sequences derived from naturally occurring transcription activator-like effectors (TALEs).[14, 15] These highly conserved 33-35 amino acids TALE repeat domains each bind to a single DNA base with binding specificity determined by the identity of two non-conserved amino acid positions.[16, 17] DNA-binding domains with novel customized specificities can be engineered by joining TALE repeats into more extended arrays. The challenge for researchers interested in utilizing TALENs is the need to construct plasmids encoding long arrays of TALE repeats that are each highly similar in sequence.

We recently described a rapid and simple method for assembling DNAs encoding extended TALE repeat arrays.[10] Our method uses a serial, hierarchical ligation strategy in which DNA fragments encoding single TALE repeats are joined together using standard restriction digest and ligation techniques that can be practiced simply and inexpensively by any laboratory. We have also developed web-based software that aids not only in identification of potential target sites within a sequence of interest but also produces a customized graphical guide illustrating the series of ligation steps required to construct each TALE repeat array. We recently demonstrated the successful use of our protocol to engineer TALENs that induced targeted mutations with high efficiencies in endogenous zebrafish genes.[10]

A variety of other platforms for constructing DNA sequences encoding TALE repeat arrays have also been described.[3, 4, 8, 11, 18-21] Nearly all of these methods rely on the use of a specialized multi-fragment ligation strategy referred to by some as Golden Gate cloning. With these approaches, DNA encoding subsets of TALE repeats (ranging in length from four- to ten-mers) are initially assembled in parallel using multi-fragment ligation reactions and then subsequently joined together to create the final desired plasmid. Because the ordered assembly of DNA fragments in these reactions relies on different sticky end overhangs for each TALE repeat at each position in an array, these methods typically require a large number of different plasmids or PCR reactions to practice. In addition, many of the methods are optimized for construction of arrays with either a fixed number or a fixed

multiple of repeats, limiting their flexibility to construct arrays of any desired length. Among all of these methods, only one provides associated software for identifying potential TALEN target sites in a sequence of interest.[8] However, this software only identifies sites and does not guide the user through the complex multi-fragment ligation process needed to assemble multiple TALENs.

Another potentially important consideration in choosing a platform for assembly is the amino acid sequence and framework of the TALE-based DNA-binding domain. A multitude of different engineered TALE frameworks have been described and utilized in the literature. Two sources of variability exist in these different frameworks: (1) the TALE repeats used differ at certain less well-conserved positions within the domain and (2) additional amino-terminal and carboxy-terminal TALE-derived sequences required for DNA-binding activity of the TALE repeat array vary in both length and amino acid sequence.[2, 3, 6] Some have noted that the choice of framework is important and can influence the activities of TALENs. We note that our platform, unlike all other publicly available methods, utilizes a framework of TALE repeats and amino-terminal and carboxy-terminal TALE-derived sequences developed by Rebar and colleagues.[3] This framework has been used successfully to construct TALENs with high activities in *C. elegans*,[9] zebrafish,[10] rats,[12] and human somatic[3] and pluripotent stem cells.[13]

Here we describe a detailed protocol for identifying potential TALEN targets within a sequence of interest and for assembling TALENs to those sites. Our assembly procedure uses standard restriction digest and ligation reactions and therefore does not require specialized expertise or performance of multi-fragment ligation reactions. Our web-based software program (freely available without registration) identifies potential TALEN target sites in user-defined sequences and generates printable, color-coded graphical guides that provide a roadmap for assembly of desired TALENs. These customized guides also give the names of specific plasmids required to assemble each particular TALEN. All plasmids required to practice our protocol are publicly available to academic researchers from non-profit plasmid distribution service Addgene (http://www.addgene.org/talengineering).

OVERVIEW OF THE PROCEDURE

Our protocol can be conceptually divided into three steps: (1) identification of potential TALEN targets within the sequence of interest; (2) assembly of plasmids encoding TALE repeat arrays; and (3) cloning of DNA fragments encoding TALE repeat arrays into a TALEN expression vector.

Identifying TALEN target sites using the ZiFiT Targeter program:

We have added a new module to our previously described Zinc Finger Targeter (ZiFiT) program that enables identification of TALEN target sites in addition to ZFN target sites. To reflect this change, we have re-named our program Zinc Finger & TALE Targeter (ZiFiT Targeter). Because TALENs can be targeted to a broad range of potential sites, ZiFiT Targeter restricts its output by default to five highest-ranked potential sites (criteria used to rank sites are based on published data describing the length of arrays and the length of spacer sequences between the TALEN binding sites as well as avoiding targeting of substantially overlapping (and therefore similar) sites; see Supplementary Discussion for chapter 5 in Appendix C). ZiFiT Targeter also provides users with customized graphical guides for assembly of each TALEN.

Assembly of plasmid DNA encoding TALE repeat arrays:

Plasmid DNAs encoding TALE repeat arrays are rapidly assembled using a serial, hierarchical assembly strategy based on simple restriction enzyme digests and standard ligations (**Figure 5.2**). The details of which particular plasmids to ligate together and in what order are provided by the graphical guide generated by ZiFiT Targeter for each TALEN. This procedure generates a plasmid encoding the final desired TALE repeat array flanked by unique restriction sites that can be used for cloning into a TALEN expression vector.

Cloning of DNA encoding TALE repeat arrays into a TALEN expression vector:

In the final step, a fragment encoding the TALE repeat array is cloned into a TALEN expression vector and sequence verified. Our TALEN expression vectors provide the final carboxy-terminal “0.5” TALE repeat domain, additional amino-terminal and carboxy-terminal TALE-derived sequences required for optimal DNA binding, and the wild-type FokI

nuclease domain. TALENs expressed from these vectors also possess a Triple FLAG epitope tag and a nuclear localization signal, both encoded at the amino-terminus. A TALEN encoded on these vectors can be expressed in cells using a CMV promoter that is present on the vector. Alternatively, the TALEN coding sequence can be transcribed into RNA *in vitro* using a T7 promoter also encoded on the expression plasmid.

MATERIALS

REAGENTS

- Plasmids encoding individual TALE repeats (available through Addgene: http://www.addgene.org/talengineering)
- TALEN Expression Vectors (available through Addgene: http://www.addgene.org/talengineering)
- Chemically competent bacterial strain XL-1 Blue (*recA1 endA1 gyrA96 thi-1 hsdR17 supE44 relA1 lac* [F´ *proAB lacIq lacZDM15* Tn*10* (Tet^R)]; Stratagene cat no. 200249)
- Carbenicillin (Sigma, cat. No C1389)
- Restriction enzymes from New England Biolabs: BamHI (cat. no. R0136S), BamHI-HF (cat. no. R3136S), BbsI (cat. no. R0539L), BsaI (cat, no. R0535L), BsmBI (cat. no. R0580L), and KpnI-HF (cat. no. R3142L)
- Bovine Serum Albumin (10 mg/ml; included with enzymes from New England Biolabs)
- NEBuffer2, NEBuffer 3, and NEBuffer 4 (10x restriction enzyme buffers included with enzymes from New England Biolabs)
- Quick Ligation Kit (New England Biolabs, cat. No. M2200L)
- QIAprep Spin Miniprep Kit (Qiagen, cat. No. 27106)
- LB medium powder (Difco, cat No. 244620)
- LB agar medium powder (Difco, cat No 244520)
- AccuGel 29:1 acrylamide:bis-acrylamide solution (National Diagnostics, cat no. EC-852)

- 10% ammonium persulfate (Fisher cat. No 7727-54-0)
- TEMED (Fisher cat. No. BP150-100)
- 100% ethanol (Pharmco, cat. No. 111ACS200)
- 70% ethanol
- Sequencing primer OK163: 5' CGCCAGGGTTTTCCCAGTCACGAC 3'
- Sequencing primer JDS2978: 5' TTGAGGCGCTGCTGACTG 3'
- Sequencing primer JDS2980: 5' TTAATTCAATATATTCATGAGGCAC 3'
- Sequencing primer JDS2778: 5' CTGGCGCAATGCGCTCAC 3'
- Sequencing primer JDS2979: 5' AAGCAATGGCGACCACCTGTTC 3'

EQUIPMENT

- 96-well PCR thermocycler
- Orbital platform shaker with adjustable speed
- Sterile bacterial culture tubes
- 1.5ml Eppendorf Tubes
- Tabletop Centrifuge

PROCEDURE

Identification of potential TALEN target sites using web-based ZiFiT Targeter software

1. Identify any repeat sequences within the target sequence of interest by entering it into the RepeatMasker Web Server (http://www.repeatmasker.org/). Repeat sequences should be excluded from any sequence that is to be analyzed for potential TALEN target sequences by ZiFiT Targeter.

2. Visit the ZiFiT Targeter website at http://zifit.partners.org/zifitbeta. (Note that this is a temporary URL for our updated version of ZiFiT Targeter – we will place this version on the main ZiFiT website at http://zifit.partners.org if our manuscript is published.)

3. Click on the ZiFiT option on the top menu and then click on the “Design TALE Nucleases” option under the “TALE Assembly” menu.

4. Paste the nucleotide sequence of interest into the text box labeled ‘Sequence’. The sequences may be entered as raw data or in FASTA format. All numbers and characters that are not G, A, T, or C will be ignored. The user must indicate the nucleotide position at which they wish the break to occur by framing it with brackets (e.g.--[A] or [G]). ZiFiT Targeter will attempt to identify sites that place this nucleotide within the spacer sequence between the TALEN target half-sites.

5. Check the option “Mask redundant sites” to make sure that the sites identified by ZiFiT Targeter are different from each other by at least 6 base pairs (3 base pairs in each TALEN monomer binding site)

6. By default, ZiFiT Targeter will only return the five highest-ranked sites it identifies using criteria described in Supplementary Discussion in appendix C. However, a user can display all the potential target sites identified by the program by checking the box for “Display all”. With this option checked, ZiFiT Targeter will return all potential TALEN target sites that consist of half-sites of lengths 12-22 base pairs (including the conserved 5’ T nucleotide) and spacer sequences of lengths 12-23 base pairs.

7. Click the Submit button. The output reports the top five TALEN target sites below the sequence entry box. If the option to “Display all” is selected, all the sites are reported on a new page.

8. To obtain a customized graphical guide for assembling a particular TALEN, click on the target site. The guide will open in a new window and can be saved or printed.

Construction of TAL Arrays

9. Using the graphical output provided by ZiFiT Targeter as a guide, perform the series of ligations indicated in the first row of the figure by using steps 10 - 19 below. Note that the numbers above each TALE repeat unit in the graphical guide identify the names of the plasmids available from Addgene that are to be used for each ligation. For example, in the graphical output shown in **Figure 5.2**, one would

perform seven ligations of the following pairs of plasmids: 7 and 14, 16 and 22, 30 and 11, 17 and 24, 26 and 15, 16 and 24, and 30 and 12). For each pair to be ligated the TALE repeat on the left is the amino-terminal repeat and the one on the right is the carboxy-terminal repeat. In each ligation, a DNA vector backbone encoding the amino-terminal repeat is ligated to a fragment encoding the carboxy-terminal repeat.

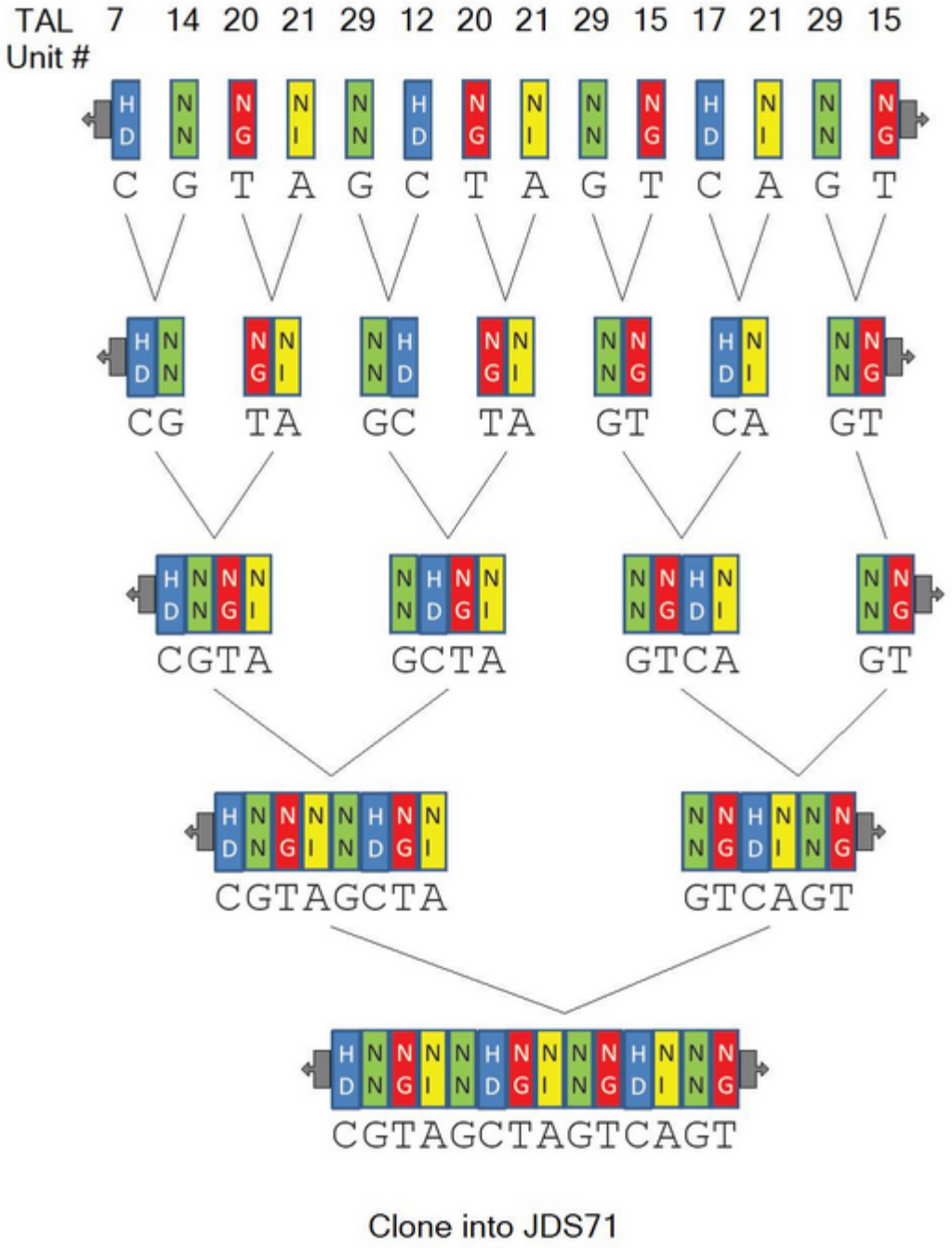


Figure 5.1. Output from ZiFiT Targeter.

10. Digest plasmid(s) encoding the amino-terminal TALE repeat(s) with BamHI and BsaI using the conditions listed below. Incubate the reaction for 2 hours at 37°C.

Component	Amount
Plasmid encoding amino-terminal TALE repeat(s)	1.5 µg
NEB Buffer 2	4 µl
Bovine Serum Albumin (10mg/ml)	4 µl
BbsI (5 U/µl)	2 µl
BamHI (20 U/µl)	2 µl
Add nuclease-free water to a total volume of	40 µl

11. Isolate the vector backbone from the restriction digest of step 10 on a 5% non-denaturing polyacrylamide gel and purify the DNA according to the protocol found in Box 2 of Maeder et al.[22] Final DNA pellets should be resuspended in 20 µl of nuclease-free water.

12. Digest plasmid(s) encoding the carboxy-terminal TALE repeat(s) with BbsI and BamHI using the conditions listed below. Incubate the reaction for 2 hours at 37°C.

Component	Amount
Plasmid encoding carboxy-terminal TALE repeat	1.5 µg
10X Buffer (NEBuffer 4)	4 µl
Bovine Serum Albumin (10mg/ml)	4 µl
BsaI (5 U/µl)	2 µl
BamHI (10 U/µl)	2 µl
Add nuclease-free water to a total volume of	40 µl

13. Isolate the DNA fragment encoding the TALE repeat(s) from the restriction digest of step 12 on a 5% non-denaturing polyacrylamide gel and purify the DNA using the protocol found in Box 2 of Maeder et al.[22] Final DNA pellets should be

resuspended in 20 µl of nuclease-free water. Note that the size of the DNA fragment will be equal to N x ~102 bps where N is the number of TALE repeats encoded on the fragment.

14. Ligate the purified DNA fragment isolated in step 13 into the purified vector backbone isolated in step 11 as tabulated below. Also perform a control ligation using vector backbone alone (i.e.--without fragment). Allow the ligation reactions to incubate for 15 minutes at room temperature.

Component	Amount
Purified vector backbone (from step 11)	1 µl
Purified fragment (from step 13) or water	3 µl
Quick Ligase Buffer (NEB)	4.5 µl
T4 DNA Ligase (400U/ul)	0.5 µl
Total	9 µl

15. Transform each ligation from step 14 into 90 µl chemically competent XL1-Blue cells. Mix ligations with competent cells and leave on ice for 5 minutes. Perform heat shock at 42°C for 1 minute then return transformations to ice for 1 minute. Add 500 µl LB and recover with agitation for 45 minutes at 37°C. Plate 200 µl of each transformation on an LB agar plate supplemented with 100 µg ml^{-1} carbenicillin and incubate overnight at 37°C for 12 – 16 hours.

16. If the actual ligation/transformation of step 15 yields at least 10-fold more colonies than the control ligation, inoculate two single colonies from the actual ligation/transformation plate into 4 ml of LB supplemented with carbenicillin 100 mg ml^{-1} and grow overnight with agitation at 37°C.

CRITICAL STEP: To reduce the risk of plasmid deletions, do not allow the cultures to grow for more than 12 hours.

17. Isolate plasmid DNA from overnight cultures using a QIAprep Spin Miniprep Kit and following the manufacturer's instructions.

18. To determine whether the plasmids isolated in step 17 have successfully taken up the fragment encoding the carboxy-terminal TALE repeat(s), digest the candidate plasmids with XbaI and BamHI as detailed below and incubate at 37°C for 1 hour.

Component	Amount
Plasmid	1 µg
10X Buffer (NEB Buffer 4)	4 µl
Bovine Serum Albumin (10mg/ml)	4 µl
XbaI (20 U/µl)	2 µl
BamHI-HF (20 U/µl)	2 µl
Add nuclease-free water to a total volume of	40 µl

19. Visualize the products of the restriction digests from step 18 on a 5% non-denaturing polyacrylamide gel. Plasmids that have successfully taken up the fragment encoding the carboxy-terminal TALE repeat(s) should yield a ([(M + N) x 102] + 33) bp fragment (where M and N are the numbers of TALE repeats encoded by the vector backbone and fragment, respectively, used for the ligation.).

20. Following completion of the first set of ligations, perform the series of ligations indicated in the second row of the graphical guide by using steps 10 - 19 above.

21. Continue performing ligations in subsequent rows of the graphical guide using steps 10 - 19 above until the final assembled array is completed.

Cloning TAL arrays into the nuclease backbone

22. Digest the specific TALEN expression vector indicated at the bottom of the graphical output (pJDS70, pJDS71, pJDS74 or pJDS78) with BsmBI restriction enzyme as detailed below. Incubate the reaction at 55°C for 3 hours.

Component	Amount
TALEN Expression Vector	2µg
10x Buffer (NEB Buffer #3)	5 µl

BsmBI (10U/ul)	5 µl
Add nuclease-free water to a total volume of	50 µl

23. Isolate the vector backbone from the restriction digest of step 22 on a 5% non-denaturing polyacrylamide gel and purify the DNA using the protocol found in Box 2 of Maeder et al.[22] Final DNA pellets should be resuspended in 20 µl of nuclease-free water.

24. Digest plasmid(s) encoding the final assembled TALE repeat array (from step 21 above) with BbsI and BsaI as tabulated below. Incubate the reaction for 2 hours at 37°C.

Component	Amount
Plasmid encoding assembled TALE repeat array	1.5 µg
10X Buffer (NEB Buffer 2)	4 µl
Bovine Serum Albumin (10mg/ml)	4 µl
BbsI (5 U/µl)	2 µl
BsaI (10 U/µl)	2 µl
Add nuclease-free water to a total volume of	40 µl

25. Isolate the DNA fragment encoding the TALE repeat array(s) from the restriction digest of step 24 on a 5% non-denaturing polyacrylamide gel and purify the DNA using the protocol found in Box 2 of Maeder et al.[22] Final DNA pellets should be resuspended in 20 µl of nuclease-free water. Note that the size of the DNA fragment will be equal to N x ~102 bps where N is the number of TALE repeats encoded on the fragment.

26. Ligate the purified DNA fragment isolated in step 25 into the purified TALEN expression vector backbone isolated in step 23 as tabulated below. Also perform a control ligation using vector backbone alone (i.e.--without fragment). Allow the ligation reactions to incubate for 15 minutes at room temperature.

Component	Amount
Purified vector backbone (from step 23)	1 µl
Purified fragment (from step 25) or water	3 µl
Quick Ligase Buffer (NEB)	4.5 µl
T4 DNA Ligase (400U/ul)	0.5 µl
Total	9 µl

27. Transform each ligation from step 26 into 90 µl chemically competent XL1-Blue cells. Mix ligations with competent cells and leave on ice for 5 minutes. Perform heat shock at 42°C for 1 minute then return transformations to ice for 1 minute. Add 500 µl LB and recover with agitation for 45 minutes at 37°C. Plate 200 µl of each transformation on an LB plate supplemented with 100 µg ml^{-1} carbenicillin and incubate overnight at 37°C for 12 – 16 hours.

28. If the actual ligation/transformation of step 27 yields at least 10-fold more colonies than the control ligation, inoculate two single colonies from the actual ligation/transformation plate into 4 ml of LB supplemented with carbenicillin 100 mg ml^{-1} and grow overnight with agitation at 37°C.

CRITICAL STEP: To reduce the risk of plasmid deletions, do not allow the cultures to grow for more than 12 hours.

29. Isolate plasmid DNA from overnight cultures using a QIAprep Spin Miniprep Kit and following the manufacturer's instructions.

30. To determine whether the TALEN expression plasmids isolated in step 29 have successfully taken up the fragment encoding the TALE repeat array(s), digest the candidate plasmids with KpnI and BamHI as detailed below and incubate at 37°C for 2 hours.

Component	Amount
Plasmid	0.5 µg
10X Buffer (NEBuffer 4)	5 µl
Bovine Serum Albumin (10 mg/ml)	5 µl

KpnI-HF (20 U/µl)	1 µl
BamHI-HF (20 U/µl)	1 µl
Add nuclease-free water to a total volume of	50 µl

31. Visualize the products of the restriction digests from step 30 on a 5% non-denaturing polyacrylamide gel. Plasmids that have successfully taken up the fragment encoding the TALE repeat array should yield a (650 + [N x ~102]) bp fragment (where N is the number of TALE repeats encoded in the array.).

32. Optional: TALEN expression plasmids can be sequence-verified by DNA sequencing using forward primer JDS2978 and reverse primer JDS2980. This step is not required because no PCR is performed during the assembly process. However, sequencing can be useful to confirm that the correct TALE repeat arrays have been ligated together in the extended array.

TALEN expression plasmids encoding 12.5 to 16.5 TALE repeats can be assembled in ~9 days. Longer arrays require an additional two days. The process of cloning the fragment encoding the assembled TALE repeat array into the TALEN expression vector requires 3 days. However, because this final cloning step is started the same day that the assembly of the TALE repeat array is completed, the assembly of a TALEN expression plasmid containing up to 16.5 TALE repeats can be completed in 11 days.

ANTICIPATED RESULTS

We have not encountered any difficulties with assembling various TALEN expression plasmids using the approach described above. A critical factor in ensuring success is to check by restriction digest analysis that the ligations work successfully at each step because one failed ligation reaction can prevent successful assembly of the final desired array. In addition, we have found that it is critical to sequence the final TALEN expression plasmid to verify that the correct TALE units have been used at each step and that the array was cloned into the correct TALEN expression vector.

AUTHOR CONTRIBUTIONS

J.D.S. and J.K.J. conceived of the assembly strategy and designed the plasmid DNAs. J.D.S., D.R., and C.K. experimentally validated and refined the strategy. J.D.S., D.R., and J.K.J. designed the ZiFiT Targeter software and wrote the paper.

ACKNOWLEDGMENTS

This work was supported by a National Institutes of Health (NIH) Director's Pioneer Award DP1 OD006862 (J.K.J.), NIH R01 GM088040 (J.K.J.), the Jim and Ann Orr MGH Research Scholar Award (J.K.J.), NIH T32 CA009216 (J.D.S.), and National Science Foundation DBI-0923827 (D.R. and J.K.J.).

Competing Financial Interests

J.D.S. and J.K.J. are inventors on a patent application describing the TALE repeat array assembly method.

REFERENCES

1. DeFrancesco, L., *Move over ZFNs.* Nat Biotechnol, 2011. **29**(8): p. 681-4.
2. Christian, M., et al., *Targeting DNA double-strand breaks with TAL effector nucleases.* Genetics, 2010. **186**(2): p. 757-61.
3. Miller, J.C., et al., *A TALE nuclease architecture for efficient genome editing.* Nat Biotechnol, 2011. **29**(2): p. 143-8.
4. Li, T., et al., *TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain.* Nucleic Acids Res, 2011. **39**(1): p. 359-72.
5. Mahfouz, M.M., et al., *De novo-engineered transcription activator-like effector (TALE) hybrid nuclease with novel DNA binding specificity creates double-strand breaks.* Proc Natl Acad Sci U S A, 2011. **108**(6): p. 2623-8.
6. Mussolino, C., et al., *A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity.* Nucleic Acids Res, 2011.
7. Li, T., et al., *Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes.* Nucleic Acids Res, 2011. **39**(14): p. 6315-25.
8. Cermak, T., et al., *Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting.* Nucleic Acids Res, 2011. **39**(12): p. e82.
9. Wood, A.J., et al., *Targeted genome editing across species using ZFNs and TALENs.* Science, 2011. **333**(6040): p. 307.

10. Sander, J.D., et al., *Targeted gene disruption in somatic zebrafish cells using engineered TALENs.* Nat Biotechnol, 2011. **29**(8): p. 697-8.
11. Huang, P., et al., *Heritable gene targeting in zebrafish using customized TALENs.* Nat Biotechnol, 2011. **29**(8): p. 699-700.
12. Tesson, L., et al., *Knockout rats generated by embryo microinjection of TALENs.* Nat Biotechnol, 2011. **29**(8): p. 695-6.
13. Hockemeyer, D., et al., *Genetic engineering of human pluripotent cells using TALE nucleases.* Nat Biotechnol, 2011. **29**(8): p. 731-4.
14. Bogdanove, A.J., S. Schornack, and T. Lahaye, *TAL effectors: finding plant genes for disease and defense.* Curr Opin Plant Biol, 2010. **13**(4): p. 394-401.
15. Scholze, H. and J. Boch, *TAL effectors are remote controls for gene activation.* Curr Opin Microbiol, 2011.
16. Moscou, M.J. and A.J. Bogdanove, *A simple cipher governs DNA recognition by TAL effectors.* Science, 2009. **326**(5959): p. 1501.
17. Boch, J., et al., *Breaking the code of DNA binding specificity of TAL-type III effectors.* Science, 2009. **326**(5959): p. 1509-12.
18. Zhang, F., et al., *Efficient construction of sequence-specific TAL effectors for modulating mammalian transcription.* Nat Biotechnol, 2011. **29**(2): p. 149-153.
19. Geissler, R., et al., *Transcriptional activators of human genes with programmable DNA-specificity.* PLoS ONE, 2011. **6**(5): p. e19509.
20. Weber, E., et al., *Assembly of designer TAL effectors by Golden Gate cloning.* PLoS ONE, 2011. **6**(5): p. e19722.
21. Morbitzer, R., et al., *Assembly of custom TALE-type DNA binding domains by modular cloning.* Nucleic Acids Res, 2011. **39**(13): p. 5790-9.
22. Maeder, M.L., et al., *Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays.* Nat Protoc, 2009. **4**(10): p. 1471-501.

CHAPTER 6. GENERAL CONCLUSIONS & FUTURE DIRECTIONS

ZFNs or TALENs – Which platform is better? As Sir Aaron Klug said, “There is only one word that matters in biology, and that is specificity.” [1] Ultimately, the more specific platform will be the better platform.

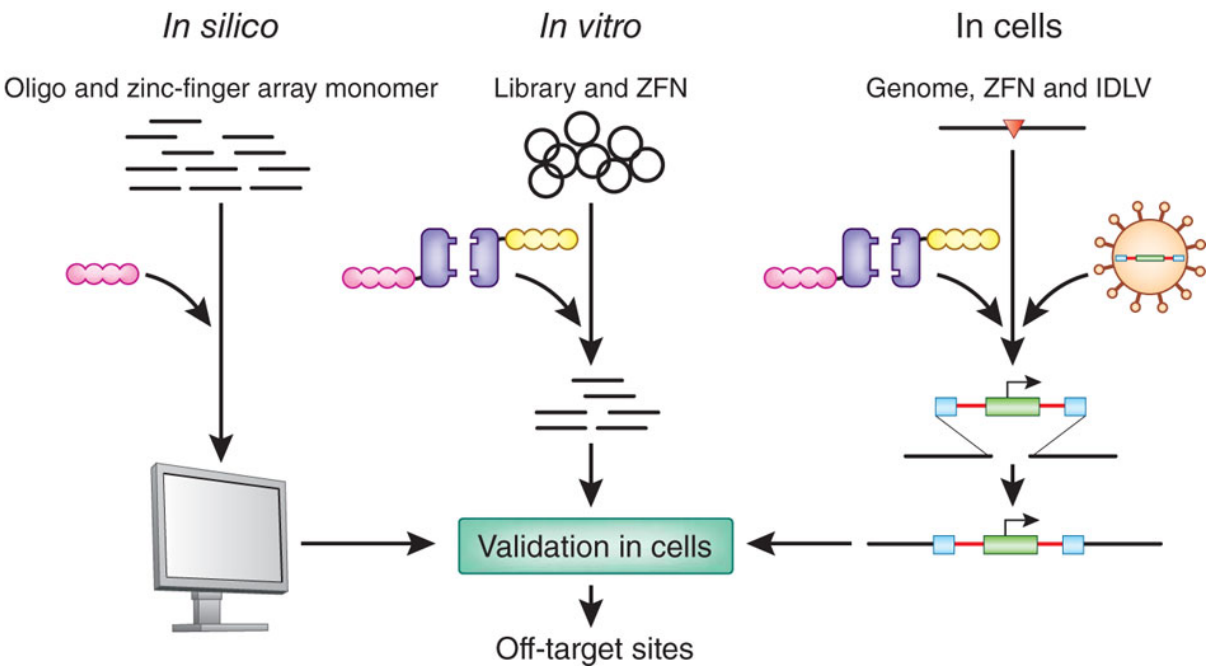


Figure 6.1. Method to determine off-target sites of ZFNs. IDLV, integrase-deficient lentivirus. (Figure from [2])

The question of which platform provides higher specificity is not easy to answer because, to draw definitive conclusions, we would need to sequence and analyze the complete genomic DNA for every cell or organism in which several ZFNs or TALENs targeted to identical sites had been tested. As this is prohibitively expensive, other more creative methods have been devised. Recently, Pattanayak et al. [3] and Gabriel et al. [4] published two different approaches (described below) to test the specificity of ZFNs in human cells. Both methods detected several off-target effects in K562 cells after introduction

of ZFNs designed by Sangamo Biosciences to target the CCR5 locus (i.e., the same ZFNs currently being tested in human clinical trials [3]).

The method described by Pattanayak et al. [3] involved the generation of a library of target sites similar to the intended target site. In theory, the library covered all DNA sequences that have seven or fewer substitutions in about 10-fold excess. The library was incubated with the ZFNs of interest and, subsequently, DNA molecules were cleaved to expose a 5' phosphate required for the ligation of sequencing primers. The library was then deep-sequenced and target sites that were preferentially cleaved were identified. Several ZFN target sites identified using this *in vitro* assay was mapped to 37 genomic loci. To check for evidence of cleavage at the corresponding sites *in vivo*, K562 cells were transfected with the same ZFNs. Of the 37 sites detected *in vitro*, 34 could be amplified, and of these 34, 10 showed evidence of modification via non-homologous end-joining (NHEJ). One of these off-target sites lies in the promoter region of a malignancy associated gene, *BTBD10.* An interesting observation in this study was the extent to which the amount of off-target cleavage was dependent on nuclease concentration. See Figure 6.2.

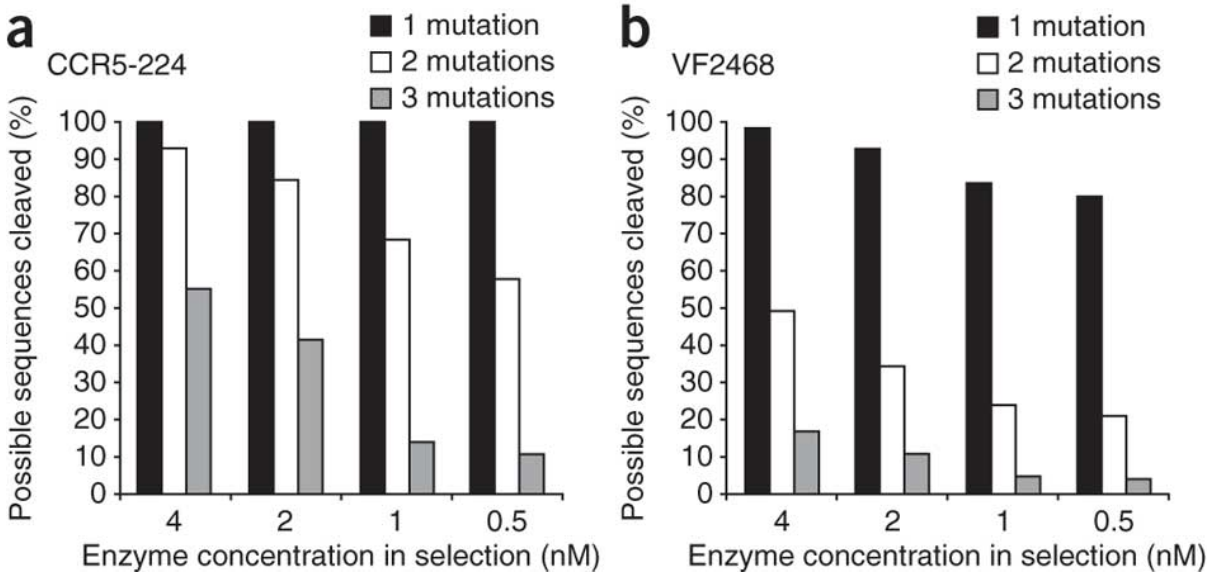


Figure 6.2 Effects of enzyme concentration on cleavage. Figure from [3]

The method described by Gabriel et al. [4] was also straight forward. It involved the integration of Integrase-Defective Lentiviral Vectors (IDLV) into the double-stranded breaks created by ZFNs. Linear Amplification Mediated (LAM) PCR was performed on ZFN-treated K562 cells that had been transfected with these vectors, using primers that hybridize to the U5 (unique 5) region on the LTR of the integrated IDLV. After filtering for noise (because IDLVs can also integrate randomly into DNA), 7 *bona fide* IDLV insertions at potential ZFN off-target sites were identified. Among these, 4 showed evidence of NHEJ-generated mutations upon deep sequencing.

Although these numbers may seem alarming at first glance, it should be noted that, from the analysis of Pattanayak et al. [3], once the CCR2 off-target site (a previously identified off-target site) is eliminated from consideration, the next most frequent off-target site is ~100-fold less likely to be cleaved; given the inherently low rate of NHEJ events (< 1% [5]) in human cells, these off-target cleavage events would indeed be extremely rare.

The specificity of TALENs haven't be analyzed is as much detail but preliminary studies are promising. The most comprehensive study to date was performed by Li et al [6], where they sequenced the genome of yeast expressing TALENs. Of the 5 strains sequenced (4 expressing TALENs and 1 expressing a ZFN) none of them showed any signs off-target effects. Some point mutations were detected, but these are unlikely to be caused by nuclease mediated NHEJ events which tend to be large deletions and insertions. In another study Hockemeyer et al. analyzed highly homologous loci for evidence of NHEJ in iPS cells transfected with TALENs designed to target the AAVS1 locus. Surprisingly, 2 of the top 20 off-target sites showed evidence of NHEJ.

We are starting to gather data about the specificity of TALENs but it is still too early to draw any concrete conclusions about how they compare to ZFNs on the bases of specificity.

CONTRIBUTIONS OF THIS DISSERTATION

In this dissertation I describe:

1. My contributions to make zinc finger nucleases more accessible as tools for genome engineering. Genome modification using ZFNs was first described by

Bibikova et al. is 2001 and was poised to revolutionize the field but stumbled because acquiring quality reagents didn't turn out to be as easy as initially thought. A schism formed in the field of ZFN design: design several ZFNs using an easy inexpensive technology (modular assembly) and find one that works; or design few ZFNs using rather cumbersome, expensive technologies (selection based) that are more likely to be functional. The debate about the superior methodology continues. My contribution to the field is to make these selection methods more appealing by helping user find ideal target sites. There that two criteria that can be exploited to identify ideal target sites: 1.) Nature of the site (i.e., composition of the target site). As described in ***Chapter 2,*** we developed a Naïve Bayes classifier to capture this information and rank order all potential target sites. 2.) Location and uniqueness of the site. Not all ZFN mediated double stranded breaks are equal. As described in ***Chapter 3***, to help researchers choose ideal sites we found all potential target sites within the genomes of several model organisms, scored them using our classifier, found how unique they were, and then displayed them within the context of other genomic features using a user-friendly genome browser.

2. My contributions to exploring the potential of newly described transcription activator like effectors (TALEs) as tools for genome engineering. A simple code for TALE-DNA bind was published in 2009 [7, 8] and appear to be ideal tools for genome engineering because them seem to be truly modular. As these proteins are new the first step is develop methods to build them easily. In ***Chapter 4,*** we describe a very simple method to build functional (in zebrafish) TALENs. In ***Chapter 5*** (under review), we describe the protocol in detail and provide a user friendly web interface to help research identify target sites, and, perhaps more importantly, guide user through the assembly process.
3. My other contributions to the field include:
    a. Developing more CoDA reagents (***Appendix A***)
    b. Upgrading ZiFiT to include CoDA reagents and ZiFOpT scores (***Appendix B***)

FUTURE STUDIES

Although the problem of designing highly sequence-specific ZFNs is far from solved, it seems likely that TALENs will dominate the spot-light in the genomic modification theater for now. For TALENs to become widely useful tools, however, there are several areas that require immediate attention: TALEN cleavage specificity and efficacy, effect of genomic context (e.g., chromatin state, epigenetic modifications), toxicity in different cell types, and effects on relative rates of homologous recombination (HR) *vs*. non-homologous end joining (NHEJ). In addition, there are many fundamental questions about TALENs that have not yet been addressed: What is their mode of DNA binding? What is their 3D structure? Because TALENs have only recently been over-expressed and purified, their kinetic and thermodynamic properties (on/off rates, binding affinity, etc.) have not yet been characterized.

Specificity: How specific are TALENS? To rigorously examine this, several different pairs of nucleases) will have to be tested using assays likes the ones described by Pattanayak et al. [3] or Gabriel et al. [4]. Although such direct experiments have not yet been performed on a large scale, we can infer a lot about the specificity of TALENs based on the published literature. We can safely conclude that TALENs are not extremely promiscuous based on data from Li et al [6], Hockemeyer et al [9], and the apparently low levels of toxicity in cells [10]. There have been mixed reports about how many mismatches a TALEN can tolerate. Based on SELEX data, Miller et al. [11] reported that a TALE monomer that binds a 17-mer could potentially tolerate up to 9 bp mismatches [11]. On the other hand, Mussolino et al. [10] reported that a *single* mistake in a 17 bp target was sufficient to abolish binding. It must be noted that the target sites that Mussolino et al, compared were the CCR5 site and CCR2 site. One crucial difference between the CCR5 site and the CCR2 site is that the CCR2 site does not have a 5’ T, which appears to be essential for TALE binding. Currently, the jury is still out about the specificity of TALENs.

Effects of genomic context: We know very little about the effect of chromatin structure on TALENs. Most of the TALENs tested to date were designed to target sites that had been previously targeted using ZFNs (hence, known to be accessible). Given that some of the most important applications of genomic modification tools would be in the context of

stem cells which often have different methylation status compared to somatic cells [12], understanding the effect of chromatin structure and genomic context on TALEN efficacy will be crucial.

Toxicity: The question of toxicity can be divided into two parts: toxicity due to lack of specificity, and inherent toxicity upon over-expression of the enzyme. For most applications to date, the inherent toxicity of either ZFNs or TALENs has not been an issue, because genome modification is typically performed at the cellular level and the introduced enzymes never encounter the immune system. But, it is conceivable that, in a clinical context, this will not be the case. It is known that ZFPs are generally non-immunogenic in humans [13], but, TALENs, given their origin (a bacterial pathogen), could potentially induce an inflammatory immune response and/or cellular toxicity.

HR vs. NHEJ rates: The relative rate of homologous and non-homologous recombination events induced by TALENs is another piece of the puzzle that we very little anything. We do know that when a ZFN cleaves DNA, it leaves a 3-bp 5’ overhang. The problem is not that straightforward when it comes to TALs, because the spacer region can vary from 10 -22 bp. We do not yet know what effect this would have on rates.

Structural Information: Since Pavleteich et al. [14] published the first structure of Zif268 in 1991, the structures of zinc finger proteins and nucleases have be scrutinized in great detail. To date, there is no structure available for a complete TALEN protein (A. Bogdanove, personal communication). The best information we have is an NMR structure of a 1.5 repeat unit of TALE PthA, combined with a SAXS profile [15].

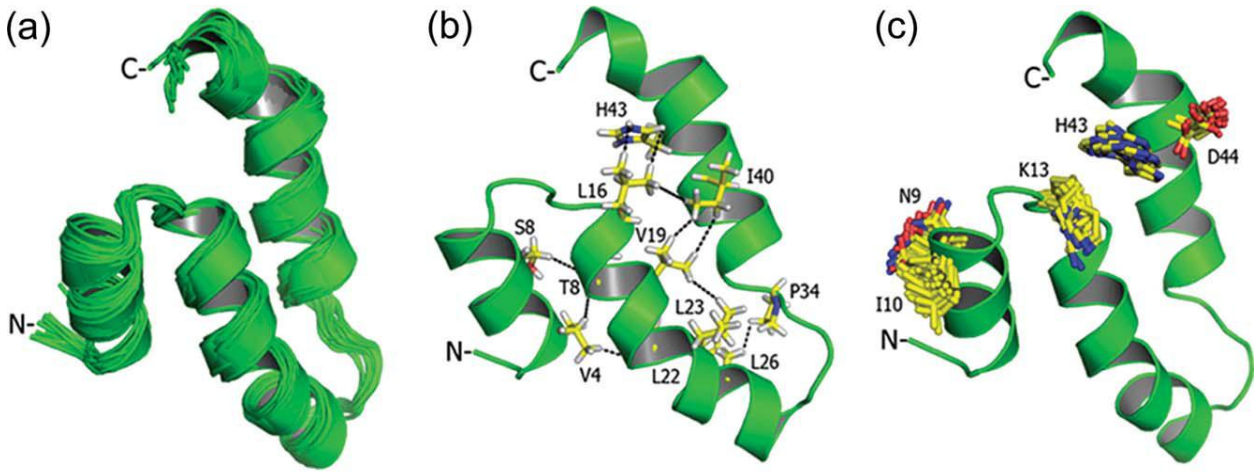
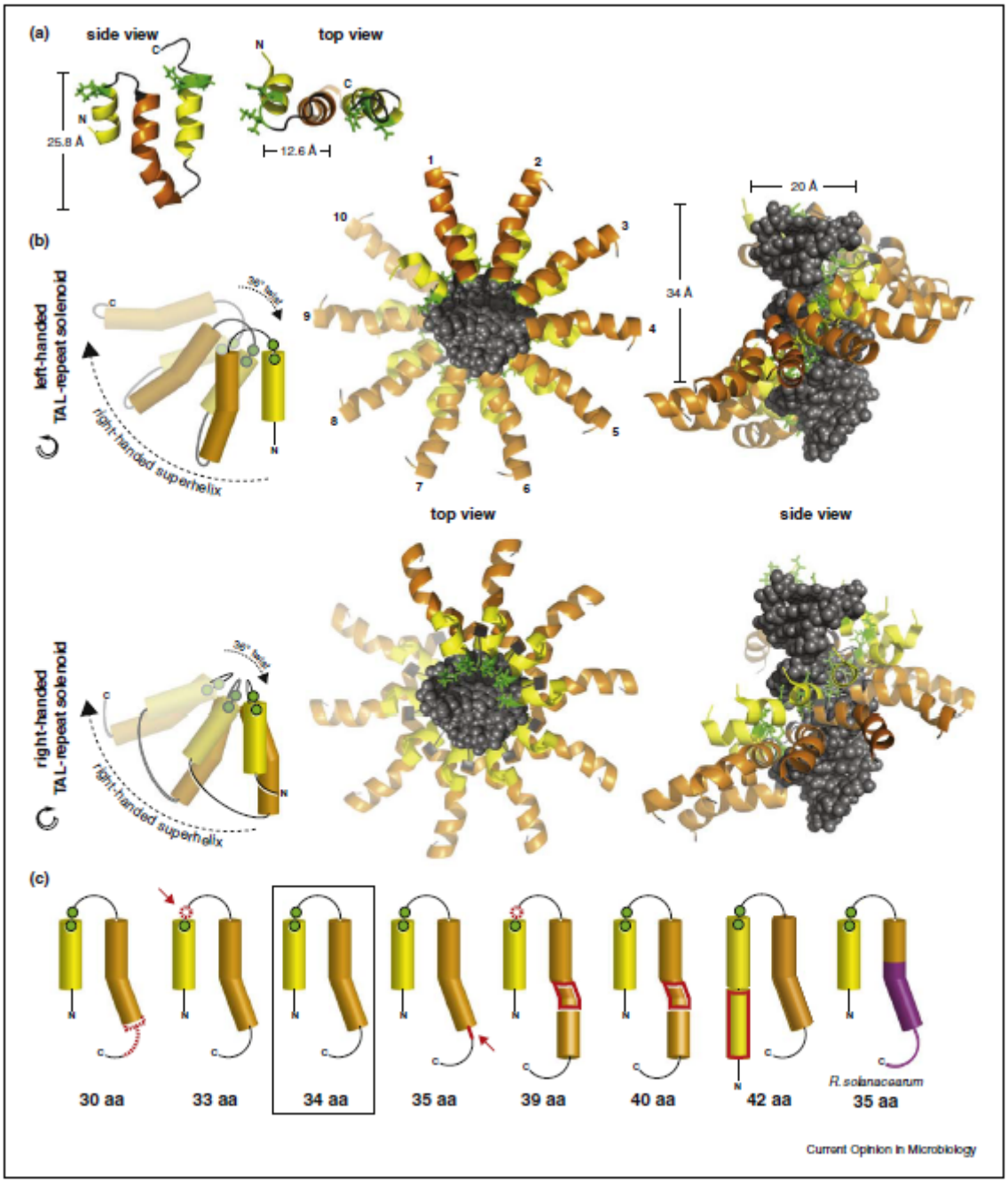


Figure 6.3. NMR structure of 1.5 repeats of the PthA TALE. (Figure from [15])

Considering the fact that a typical TALE contains ~17.5 repeats, together with a 200-250 amino acid extensions on both the N- and C-termini, it is difficult to draw too many conclusions from this structure of the 1.5 repeats. Nevertheless, these data, combined with the published SAXS profile, have been used to make predictions of the overall structure of a TALEN and its complex with DNA [13].



Possible TAL effector repeat arrangements. **(a)** NMR structure of 1.5 repeats of the TAL effector PthA2 (2KQ5) forming three α-helices [23••]. **(b)** Cartoon of three consecutive repeats forming a predicted right-handed superhelix of repeats twisted 36° to fit the right-handed DNA helix. Manual fit of 10 individual repeats onto DNA (black). To connect individual helices, the long alpha helices would have to be kinked. Top and bottom models are based on left-handed and right-handed solenoids (repeat twist), respectively. Repeats of the right-handed solenoids are shown tilted. α-Helical

Figure 6.4. Predicted structure of PthA2 with DNA. (Figure from [13])

Understanding the structure of TALENs and the mechanism by which they bind DNA is arguably the most crucial missing piece of the puzzle at the moment.

REFERENCES

1. Pearson, H., *Protein engineering: The fate of fingers.* Nature, 2008. **455**(7210): p. 160-4.
2. Mussolino, C. and T. Cathomen, *On target? Tracing zinc-finger-nuclease specificity.* Nat Methods, 2011. **8**(9): p. 725-6.
3. Pattanayak, V., et al., *Revealing off-target cleavage specificities of zinc-finger nucleases by in vitro selection.* Nat Methods, 2011. **8**(9): p. 765-70.
4. Gabriel, R., et al., *An unbiased genome-wide analysis of zinc-finger nuclease specificity.* Nat Biotechnol, 2011. **29**(9): p. 816-23.
5. Zou, J., et al., *Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells.* Cell Stem Cell, 2009. **5**(1): p. 97-110.
6. Li, T., et al., *Modularly assembled designer TAL effector nucleases for targeted gene knockout and gene replacement in eukaryotes.* Nucleic Acids Res, 2011. **39**(14): p. 6315-25.
7. Boch, J., et al., *Breaking the code of DNA binding specificity of TAL-type III effectors.* Science, 2009. **326**(5959): p. 1509-12.
8. Moscou, M.J. and A.J. Bogdanove, *A simple cipher governs DNA recognition by TAL effectors.* Science, 2009. **326**(5959): p. 1501.
9. Hockemeyer, D., et al., *Genetic engineering of human pluripotent cells using TALE nucleases.* Nat Biotechnol, 2011. **29**(8): p. 731-4.
10. Mussolino, C., et al., *A novel TALE nuclease scaffold enables high genome editing activity in combination with low toxicity.* Nucleic Acids Res, 2011.
11. Miller, J.C., et al., *A TALE nuclease architecture for efficient genome editing.* Nat Biotechnol, 2011. **29**(2): p. 143-8.
12. Han, J.W. and Y.S. Yoon, *Epigenetic Landscape of Pluripotent Stem Cells.* Antioxid Redox Signal, 2011.
13. Scholze, H. and J. Boch, *TAL effectors are remote controls for gene activation.* Curr Opin Microbiol, 2011. **14**(1): p. 47-53.
14. Pavletich, N.P. and C.O. Pabo, *Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A.* Science, 1991. **252**(5007): p. 809-17.
15. Murakami, M.T., et al., *The repeat domain of the type III effector protein PthA shows a TPR-like structure and undergoes conformational changes upon DNA interaction.* Proteins, 2010. **78**(16): p. 3386-95.

APPENDIX A. SELECTION-FREE ZINC-FINGER-NUCLEASE ENGINEERING BY CONTEXT-DEPENDENT ASSEMBLY (CODA)

Jeffry D. Sander, Elizabeth J. Dahlborg, Mathew J. Goodwin, Lindsay Cade, Feng Zhang, Daniel Cifuentes, Shaun J. Curtin, Jessica S. Blackburn, Stacey Thibodeau-Beganny, Yiping Qi, Christopher J. Pierick, Ellen Hoffman, Morgan L. Maeder, Cyd Khayter, Deepak Reyon, Drena Dobbs, David M. Langenau, Robert M. Stupar, Antonio J. Giraldez, Daniel F. Voytas, Randall T. Peterson, Jing-Ruey J. Yeh, and J. Keith Joung. *Nat Methods*., 2010 December 12.

ABSTRACT

Engineered zinc-finger nucleases (ZFNs) enable targeted genome modification. Here we describe Context-Dependent Assembly (CoDA), a platform for engineering ZFNs using only standard cloning techniques or custom DNA synthesis. Using CoDA ZFNs, we rapidly altered 20 genes in zebrafish, Arabidopsis, and soybean. The simplicity and efficacy of CoDA will enable broad adoption of ZFN technology and make possible large-scale projects focused on multi-gene pathways or genome-wide alterations.

MAIN

Engineered zinc-finger nucleases (ZFNs) can be used to introduce targeted alterations into genomes of model organisms, plants and human cells [1, 2]. Repair of ZFN-induced double-strand breaks by error-prone nonhomologous end joining leads to efficient introduction of insertion or deletion mutations at the site of the double-strand break. Alternatively, repair of a double-strand break by homology-directed repair with an exogenously introduced donor template can promote efficient introduction of alterations or insertions at or near the break site.

Widespread adoption and large-scale use of ZFN technology have been hindered by continued lack of a robust, easy-to-use and publicly available method for engineering zinc-finger arrays. In one approach, known as modular assembly, preselected zinc-finger modules are joined into arrays [3], a procedure simple enough for any researcher to implement. Some recent reports have indicated a high failure rate for this method [4, 5], although the

consequent need to construct and test large numbers of ZFNs for any given target gene can be mitigated by using a more limited subset of modules [6]. We recently described a robust selection-based method known as oligomerized pool engineering (OPEN) [7], but the labor and expertise required to screen combinatorial libraries have limited its broad adoption [3]. Researchers at Sangamo BioSciences have also developed a platform for engineering ZFNs; although some details of this method have been published [8], implementation requires access to a proprietary archive of engineered zinc-finger units [9]. Researchers may purchase customized ZFNs made by the Sangamo approach through the Sigma-Aldrich CompoZr service, but the cost of these proteins [9] limits the scale and scope of projects that can be performed.

Here we describe context-dependent assembly (CoDA), a publicly available platform of reagents and software that is simple to practice and has a success rate for generating active zinc-finger arrays comparable to that of selection-based methods such as OPEN. With the CoDA approach, three-finger arrays are assembled using N- and C-terminal fingers that have been previously identified in other arrays containing a common middle finger (F2 units) (Fig. 1). CoDA can be implemented by using a large archive consisting of 319 N-terminal-end fingers (F1 units) and 344 C-terminal-end fingers (F3 units) (Supplementary Tables 1 and 2) engineered to function well when positioned adjacent to one of 18 fixed F2 units (Online Methods). Thus, in contrast to modular assembly, CoDA does not treat fingers as independent modules but instead explicitly accounts for context-dependent effects between adjacent fingers [10, 11], thereby increasing the probability that a multifinger array will function well. CoDA is rapid and requires neither specialized expertise nor labor-intensive selections; dozens of multifinger arrays can be constructed in 1–2 weeks or less using standard cloning techniques or commercial DNA synthesis.

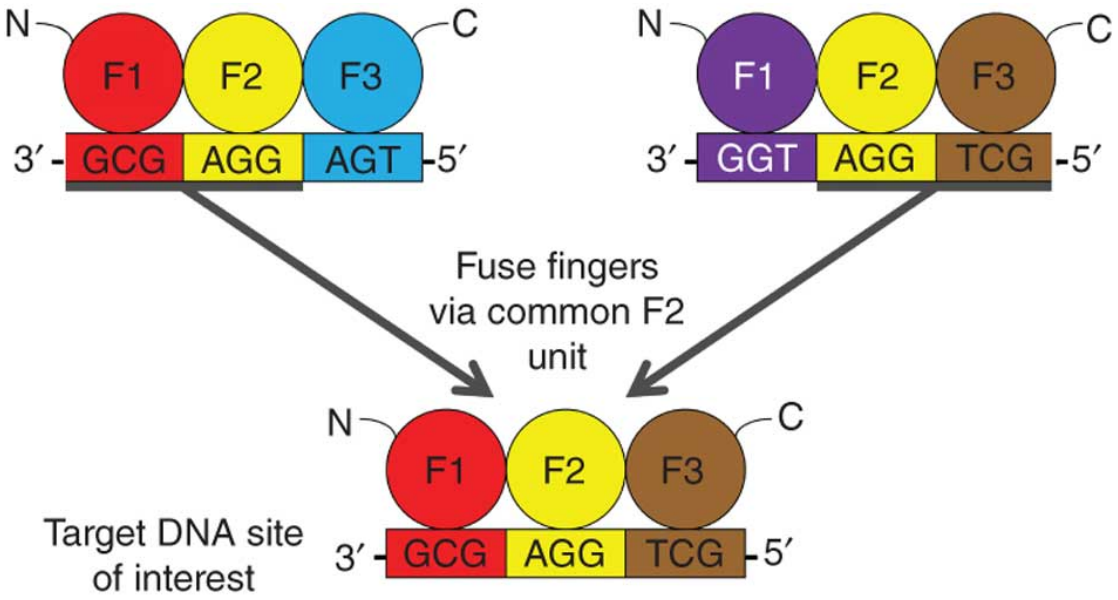


Figure. A. 1: Schematic overview of Context-Dependent Assembly (CoDA). Zinc fingers (units F1–F3) and their respective 3-bp DNA 'subsites' are shown. Two different three-zinc-finger arrays, each engineered to bind different 9-bp target sites and that have in common a middle F2 unit, can be used to create a three-finger array with a new specificity by joining together the F1 unit from the first array, the F2 unit, and the F3 unit from the second array.

To test the CoDA approach, we assembled 181 three-finger arrays and evaluated each for its ability to bind its cognate DNA target site using an established bacterial two-hybrid (B2H) reporter assay [5, 7]. Previous work has shown that three-finger arrays that do not activate transcription by more than 1.57-fold in the B2H reporter assay are likely to be inactive as ZFNs in mammalian cells [5] and those that activate by threefold or more have a high probability of functioning efficiently as ZFNs in zebrafish, plant and human cells [7, 12-15]. Of the 181 CoDA-generated arrays we tested using the B2H reporter assay, <8% (14 arrays) activated transcription by <1.57-fold and >76% (139 arrays) activated transcription by greater than threefold (Supplementary Fig. 1 and Supplementary Table 3). These 'failure' and 'success' rates for DNA-binding activity (as predicted by the B2H reporter assay) are

comparable to what we have previously observed with three-finger arrays made by OPEN [7]. Because so few (<25%) of the CoDA-generated arrays we tested gave less than threefold activation in the B2H reporter assay, our results suggest that one could skip the B2H reporter assay step and instead directly test the arrays in the cell type of interest.

We compared the efficacy of CoDA with that of modular assembly by using both approaches to construct three-finger arrays for 26 different 9–base-pair (bp) sites and by testing these proteins for DNA-binding activity in the B2H reporter assay (Supplementary Table 4). For these sites that we could target by both methods, CoDA-generated zinc-finger arrays performed better than those generated by modular assembly as judged by multiple comparisons of their DNA-binding activities (Supplementary Fig. 2 and Supplementary Discussion). The most likely explanation for the relatively higher rates of generating active arrays by CoDA is its explicit consideration of context-dependent activities between fingers [10, 11]. This difference will be more pronounced when one targets a ZFN site because two functional arrays must be engineered to create a ZFN dimer.

We used CoDA to engineer ZFNs for endogenous gene targets in zebrafish and plants. Using CoDA-generated zinc-finger arrays that activated transcription at least threefold in the B2H reporter assay, we constructed ZFN pairs for 24 gene targets in zebrafish, 13 gene targets in *Arabidopsis thaliana* and one target present in two duplicated genes in soybean (Table 1). CoDA-generated ZFNs induced targeted insertion or deletion mutations with high efficiencies in 12 of 24 zebrafish target sites (mutation frequencies, ≤1% to 16.7%), in 6 of 13*Arabidopsis* gene targets (mutation frequencies, ≤1% to 8.4%) and in a target site present in two duplicated soybean genes in transformed root tissue (mutation frequencies, 18.8% and 10.7%) (Table 1 and Supplementary Figs. 3 and 4).

Table A.1: Endogenous zebrafish and plant genes targeted by CoDA-generated ZFNs

Gene symbol	Organism	ZFN target site	Mutant alleles/total alleles	Mutation frequency (%)
[a]Target sites in each gene are written 5′ to 3′ with the two half-sites targeted by the zinc-finger arrays (uppercase letters) and the intervening spacer sequence (lowercase letters).				
Dcl4a	Soybean	TGCTTCATCacaatGGAGATGAT	6/32	18.8
Dcl4b	Soybean	TGCTTCATCacaatGGAGATGAT	3/28	10.7
MPK8	Thale cress	CTCCACAACatcagGATGACGAA	7/83	8.4
MPK11	Thale cress	CTCTTCGTCctatcgGCAGAGGCG	3/90	3.3
MKK9	Thale cress	GCCAGCGACggtggtGGTGGTGGC	3/95	3.2
MPK15	Thale cress	TTCTTCATCcagatGTTGTTGAG	2/73	2.7
MAPKKK18	Thale cress	CCCTTCCACaacaacGGAGAAGCT	2/75	2.7
GA3OX2	Thale cress	AGCTACGCCgtagccGGAGACGCC	1/94	≤1
MAPKKK1	Thale cress	GGCACCTCCgatttcGTGGAGGAA	0/190	0
MAPKKK12	Thale cress	TCCTCCACCgaatcGACGGCGCT	0/187	0
MAPKKK12	Thale cress	TTCCTCCACCgaatcGACGGCGCT	0/186	0
MAPKKK4	Thale cress	GTCTCCGCCtaggaGATGCAGAC	0/190	0
MPK15	Thale cress	TGCTTCTTCatccaGATGTTGTT	0/94	0
MPK4	Thale cress	CTCTTCGTCctatcgGTAGAGGCG	0/190	0
TZP	Thale	TTCGTCTTCgagtcGTCGTTGTT	0/141	0

Gene symbol	Organism	ZFN target site	Mutant alleles/total alleles	Mutation frequency (%)
	cress			
actn1	Zebrafish	GCCTTCTCCggggcGCAGAAGGT	10/60	16.7
rag2	Zebrafish	ATCTTCTGCtccaggGGTGAAGGT	4/52	7.7
gad2	Zebrafish	AGCCGCAGCtctcgGCTGTAGAC	3/43	7
lmna	Zebrafish	CTCTTCTCCcccagaGCTGTGGAG	2/41	4.9
apoeb	Zebrafish	CCCCTCAGCccagaTGGGAGGAG	3/64	4.7
trpm7	Zebrafish	CACACCTGCacacaGATGCTGCT	2/55	3.6
grip1	Zebrafish	GGCCACCTCcaccaGCAGCGGGC	3/90	3.3
pclo	Zebrafish	CCCCTCTCCtcaaaGCAGATGCA	3/96	3.1
jak3	Zebrafish	GGCCCCACCaagcctGCTGGAGGA	1/71	≤1
ago1	Zebrafish	CTCTGCCGCcacctaGAGGATGGT	1/96	≤1
slitrk1	Zebrafish	GCCCACAGCaatggcGGAGCCGCC	1/96	≤1
bmpr2a	Zebrafish	GACTTCCTCtctgtGCAGTCGGC	1/117	≤1
bmpr2a	Zebrafish	ACCTCCTGCagtgtGAGGTTGTC	0/156	0
cnot1	Zebrafish	GGCGTCCACgtacgaGCGGAGGAG	0/93	0
ctcf	Zebrafish	TTCCTCCTCctgatGCGGAGGCT	0/96	0
dicer1	Zebrafish	TTCTGCAGCtcaatGGAGATGGT	0/96	0
dicer1	Zebrafish	AGCTTCCTCcgccgGAAGTTGAG	0/96	0
drosha	Zebrafish	GTCCTCCTCatggcgGTCGATGGT	0/96	0
g6pcb	Zebrafish	TCCCACTGCtgattGTAGGTGGA	0/134	0
nedd4l	Zebrafish	AACCGCACCacacaGTGGAAGAG	0/86	0
nod2	Zebrafish	AACTACAACattaggGCTGGAGGA	0/103	0
rag1	Zebrafish	GTCCTCCCCttcaaGTCGAATAG	0/91	0
th2	Zebrafish	CTCCTCCTCaaacacGAAGCTGTC	0/142	0
tp53	Zebrafish	AGCAGCTGCatgggGGGGATGAA	0/107	0

Our overall per-target success rate for obtaining mutations with CoDA-generated ZFNs is 50% (19 of 38 target sites) in zebrafish and plants, a frequency comparable to our success rates of ~67% (16 of 24 target sites) with OPEN-generated ZFNs in zebrafish, plants and human cells (refs. [7, 12-15] and unpublished data). For CoDA, success rates for obtaining mutations as calculated per ZFN pair and per ZFN target site are the same because only a single CoDA-generated ZFN pair is tested per ZFN target site. Although we do not know why some CoDA- and OPEN-generated ZFNs do not induce mutations, we hypothesize that chromatin state or DNA methylation of the site, or protein stability or folding might be responsible. Regardless of the precise mechanism, we recommend that users of CoDA plan to make ZFNs for at least two target sites per gene of interest to increase the likelihood that at least one pair will introduce mutations.

CoDA still has some limitations compared to existing methods. Although modular assembly was less efficient than CoDA in our direct comparisons, modular assembly can potentially be used to target sites that CoDA currently cannot target [4, 6], and one recent report demonstrated a comparable success rate of 23% for modular assembly using a more limited subset of modules [6]. In addition, although CoDA accounts for context-dependence between adjacent fingers, it also has some limitations relative to selection-based methods such as OPEN. For example, CoDA constrains the identity of the F2 unit and does not 'balance' the effects of all three fingers on affinity and specificity of the final array. In addition, CoDA in its current form guides assembly of arrays to 9-bp target sites, ignoring the identities of the adjacent upstream and downstream bases. Thus, for highly demanding therapeutic applications (for example, introduction of alterations into human pluripotent stem cells [15]), ZFNs made by OPEN may still be preferable to those made by CoDA, and it may be necessary to engineer zinc-finger arrays with greater specificities. Nonetheless, our overall results demonstrate that CoDA is a method for assembling zinc-finger arrays that accounts for context-dependent effects, is easier to perform than OPEN selections and yields ZFNs that efficiently modify genes.

With the current archive of CoDA units, a potential ZFN target site can be found approximately once in every 500 bp of random sequence (Supplementary Discussion). However, actual targeting range can be higher, depending on genomic sequences. For

example, ~81% of 27,305 unique protein-coding transcripts in the zebrafish genome (Ensembl Zv8.57 database) contain one or more potentially targetable ZFN sites (mean, 4.37 sites), a frequency equivalent to one potential site every ~400 bp of transcript-coding sequence. By contrast, ~63% of 33,200 unique protein coding transcripts in the *Arabidopsis* genome (The Arabidopsis Information Resource 9 release) contain one or more potential ZFN target sites (mean, 2.45 sites), a frequency equal to one potential site every ~790 bp of transcript-coding sequence. We updated our publicly available web-based zinc finger targeter (ZiFiT) program (http://bindr.gdcb.iastate.edu/ZiFiT/ or http://www.zincfingers.org/software-tools.htm) to enable users to identify potential CoDA ZFN target sites in any given gene sequence (Supplementary Fig. 5 and Supplementary Discussion).

In summary, CoDA is an effective alternative method for using publicly available reagents to engineer ZFNs. The rapidity and high success rate of CoDA enabled us to mutate 20 endogenous genes in three different organisms. CoDA will foster broader adoption of ZFN technology and also enable large-scale ZFN projects focused on multigene pathways or genome-wide alterations that are difficult to implement using existing methodologies.

METHODS

Identification of finger units for practicing CoDA

To identify 'fixed' F2 fingers for various 3-bp target subsites, we analyzed the amino acid sequences of F2 units from a collection of three-finger arrays previously identified from OPEN selections performed for over 130 different 9-bp sites (refs. [7, 12-15] and M. Maeder *et al*., unpublished data). From this analysis, we identified F2 units for 18 different 3-bp subsites that occurred in at least two or more different contexts. The F1 and F3 units found adjacent to these F2 units were also chosen as CoDA units because they had been selected to work well together. To obtain additional F1 and F3 CoDA units for other 3-bp subsites, we performed OPEN selections in which we interrogated combinatorial three-finger array libraries composed of a fixed F2 unit and randomized F1 and F3 units for binding to specific 9-bp target sequences. From these selections, we analyzed the amino acid sequences of three-finger arrays that activated transcription threefold or more in the B2H reporter assay

to identify additional F1 and F3 finger units that worked well when positioned next to a specific fixed F2 CoDA unit. For selections that yielded multiple three-finger array clones, we chose F1 and F3 units that occurred the most frequently in multiple distinct arrays and/or that were found in three-finger arrays that gave the highest-fold activation in the B2H reporter assay. OPEN selections were performed essentially as described previously [7, 16] but with the modification that a beta-lactamase antibiotic resistance gene was used for selection instead of the *HIS3* gene. This modified version of OPEN enabled selections to be performed with higher throughput (M.J.G. *et al.*, unpublished data). Each of the three-finger arrays from which the F1, F2 and F3 units were derived was determined to be active in a B2H reporter assay.

Construction of zinc-finger arrays by modular assembly

Construction of plasmids encoding the modularly assembled zinc-finger arrays used in this study has been described previously [5].

Construction of zinc-finger arrays by CoDA

To assemble CoDA zinc-finger arrays, DNA fragments encoding an F1–F2 cassette or an F3 cassette were amplified by PCR from plasmids using primer pairs OK1424 and OK1427 or OK1428 and OK1429, respectively. (Sequences of all primers are listed in Supplementary Table 5). The resulting PCR products were digested with DpnI (New England Biolabs) to degrade template plasmid DNA and purified using a Qiagen PCR purification kit. The cassettes were then fused together and amplified in a single PCR step using primers OK1430 and OK1432. PCR products encoding a three-finger array were then purified using a Qiagen PCR purification kit, treated with Pfu polymerase (Stratagene) in the presence of dTTP nucleotide to create overhangs, phosphorylated with T4 polynucleotide kinase (New England Biolabs) and ligated to a B2H expression plasmid (pMG414) in which the zinc-finger array is expressed as a fusion to a fragment of the yeast Gal11P protein [16]. All plasmids were sequenced using primer OK61.

B2H reporter assay

Zinc-finger arrays made by modular assembly or CoDA were each tested for binding to its cognate target site by measuring its ability to activate transcription in the B2H reporter assay as described previously [16, 17]. All assays were performed in triplicate.

Zebrafish gene mutation analysis

Injection of zebrafish embryos, isolation of genomic DNA, limited-cycle PCR amplification of the locus of interest, cloning of PCR fragments using the TOPO TA Cloning kit (Invitrogen) and transformation of *Escherichia coli* were performed as described previously [12, 18]. Resulting colonies were assessed for gene mutations by one of two methods: (i) direct sequencing of individual clones or (ii) screening of three pooled clones for alterations in PCR fragment size using fluorescence-based analysis as described previously [18], followed by identification of specific mutations by direct sequencing.

Arabidopsis gene mutation analysis

ZFN transgene expression constructs, *Arabidopsis* transformation methods, induction of ZFN expression in *Arabidopsis* seedlings by β-estradiol and isolation of *Arabidopsis* genomic DNA were done as described previously [14]. ZFN recognition sites in the *Arabidopsis* genomic DNA were amplified by PCR, the resulting fragments were cloned using the TOPO TA Cloning kit, and DNA from individual colonies was sequenced to identify mutations at the ZFN recognition site.

Soybean gene mutation analysis

Cotyledons of the soybean variety Bert were transformed using a previously described hairy root transformation protocol [19]. The ZFN transgene was induced by application of 10 μM of β-estradiol (Sigma) on tissue culture medium. Hairy root DNA was isolated using the Qiagen DNeasy kit. Transformed roots were screened for the ZFN transgene using primers (forward primer, 5′-TGGATATGTATATGGTGGTAATGC-3′ and reverse primer, 5′-TTGAGCTTGTGGCGCAGCTCG-3′). Roots containing the transgene were then screened for mutations by a cleaved amplified polymorphic sequence (CAPS) analysis (forward primer, 5′-GTAAAAGATGTTGAAAGAAAGTTGG-3′ and reverse

primer, 5′-GCTTTTGACTTGAGCATGATGG-3′) using restriction enzyme MslI, which digests the nucleotide sequence targeted for mutagenesis. A single root was identified as carrying putative mutations in the *Dcl4a* and *Dcl4b* genes. The targeted regions of *Dcl4a* and*Dcl4b* were amplified by PCR from this root using the CAPS primers. PCR fragments were cloned in pGem T-easy (Promega) and colony PCR products for 60 clones were subsequently sequenced. Mutations were identified via sequence alignments using MEGA 4.1 (ref. [20]).

Identification of potential CoDA ZFN sites in *D. rerio* and *Arabidopsis*

Potential ZFN target sites in *D. rerio* and *Arabidopsis* were identified from the Ensembl (Zv8.57) and The *Arabidopsis* information resource (TAIR9) chromosomal assemblies and gene table files. Potential ZFN target sites were defined as those that could be targeted using the CoDA reagents described here and that had a spacer sequence of 5, 6 or 7 nucleotides that was entirely within an exon.

ACKNOWLEDGEMENTS

This work was supported by National Institutes of Health grants R01 GM088040 (J.K.J. & R.T.P.), R01 GM069906 (J.K.J.), R24 GM078369 (J.K.J.), R01 CA140188 (J-R.J.Y.), R01 GM081602 (A.J.G.), RC2 MH089956 (A.J.G.), K01 AG031300 (J-R.J.Y), K01 AR055619 (D.M.L.), T32 CA009216 (J.D.S. & J.S.B.); by National Science Foundation grant DBI 0923827 (D.F.V., D.D., & J.K.J); by the Claflin Distinguished Scholar Award (J-R.J.Y.); by the Minnesota Soybean Research and Promotion Council (S.J.C. & R.M.S.), and by Alex’s Lemonade Stand and the Leukemia Research Foundation (D.M.L.).

REFERENCES

1. Carroll, D., Gene Ther., 2008. **15**: p. 1463-1468.
2. Cathomen, T. and J.K. Joung, Mol. Ther., 2008. **16**: p. 1200-1207.
3. Kim, J.S., H.J. Lee, and D. Carroll, Nat. Methods, 2010. **7**: p. 91.
4. Kim, H.J., et al., Genome Res., 2009. **19**: p. 1279-1288.
5. Ramirez, C.L., Nat. Methods, 2008. **5**: p. 374-375.
6. Lee, H.J., E. Kim, and J.S. Kim, Genome Res., 2010. **20**: p. 81-89.
7. Maeder, M.L., Mol. Cell, 2008. **31**: p. 294-301.
8. Doyon, Y., Nat. Biotechnol., 2008. **26**: p. 702-708.

9. Pearson, H., Nature, 2008. **455**: p. 160-164.
10. Isalan, M., Y. Choo, and A. Klug, Proc. Natl. Acad. Sci. USA, 1997. **94**: p. 5617-5621.
11. Isalan, M., A. Klug, and Y. Choo, Biochemistry, 1998. **37**: p. 12026-12033.
12. Foley, J.E., PLoS ONE, 2009. **4**: p. e4348.
13. Townsend, J.A., Nature, 2009. **459**: p. 442-445.
14. Zhang, F., Proc. Natl. Acad. Sci. USA, 2010. **107**: p. 12028-12033.
15. Zou, J., Cell Stem Cell, 2009. **5**: p. 97-110.
16. Maeder, M.L., et al., Nat. Protoc., 2009. **4**: p. 1471-1501.
17. Wright, D.A., Nat. Protoc., 2006. **1**: p. 1637-1652.
18. Foley, J.E., Nat. Protoc., 2009. **4**: p. 1855-1867.
19. Govindarajulu, M., et al., Mol. Plant Microbe Interact., 2008. **21**: p. 1027-1035.
20. Tamura, K., et al., Mol. Biol. Evol., 2007. **24**: p. 1596-1599.

SUPPLEMENTARY DISCUSSION

Direct comparisons of CoDA and modular assembly zinc finger arrays for 26 target DNA sites

The DNA sites used for this experiment were chosen from among 104 sites we had previously tested to assess the efficacy of modular assembly [6] (Supplementary Table 4) and represent all of these sites for which finger arrays can currently be made using CoDA. Nearly all of these sites (24 out of 26) matched the consensus sequence 5'GNNGNNGNN3', a category of target sites for which modular assembly showed the highest success rates in our earlier report [6]. In addition, it is important to note that although we made and tested only one CoDA finger array for each of the 26 target sites, multiple modularly assembled arrays (two to six arrays) were made and tested for nearly all (25 of the 26) sites (Supplementary Table 4), using three previously published module archives.[2-6]

Our results demonstrate that CoDA yielded the zinc finger array with the highest B2H assay activity for 20 of the 26 target sites (Supplementary Table 4). Furthermore, the mean B2H fold activation of all CoDA proteins tested (5.59-fold) is higher than those made using the three different modular assembly sets (1.43-, 2.11-, and 2.53-fold; Supplementary Table 4). To further compare CoDA and modular assembly, we examined fold-activation values in the B2H reporter assay of the most active protein made by each of the two methods for the 26 target DNA sites. Of these proteins, ~38% of the modular assembled arrays activated transcription by 1.57 fold or less in the B2H compared with 0% of the CoDA arrays (Supplementary Figure 2). Furthermore, ~23% of the modularly assembled arrays activated transcription by three-fold or more in the B2H assay compared with ~69% of the CoDA arrays (Supplementary Figure 2).

Comparison of mutation frequencies induced by ZFNs made using CoDA and other engineering platforms

The range of CoDA ZFN-induced mutation frequencies we observed in zebrafish somatic cell experiments (≤1% to 16.7%) are similar but somewhat lower than those from previously published experiments. Somatic mutation rates were reported to be 3% to 20%

for ZFNs made by OPEN or B1H selection [7, 8] and 3% to 32% for ZFNs made by the proprietary Sangamo

platform.[9] Nonetheless, our previous experience suggests that the frequencies of somatic mutations we observed are high enough to make it likely that germline founders could be readily identified using these ZFNs (ref. 7 and unpublished data). For Arabidopsis, the frequencies of mutagenesis (as measured by number of mutated alleles) induced by CoDA ZFNs are comparable to those previously observed with ZFNs made by OPEN.[10] No comparisons to prior experiments could be made for the soybean experiments because, to our knowledge, these are the first examples of ZFN- targeted mutations in endogenous soybean genes.

Predicted Targeting Range of CoDA ZFNs in Random DNA Sequence

The number of 9 bp sequences that can be targeted for each F2 triplet can be calculated as the product of F1 and F3 domains selected to function well in the context of the fixed F2 anchor finger for that triplet (i.e. CoDA targets where F2 target is GGG = 23 F1s x 20 F3s). Therefore, total number of 9 bp sequences that can be targeted with our current CoDA finger units is the sum of the 9bp targets for each of the 18 different F2 triplets which equals 6680 9bp sequences or approximately 2.55% of all possible 9bp DNA sequences. Because each ZFN requires two 9bp arrays and can be designed with three spacer sizes (5, 6 or 7 bp), [11, 12] CoDA targets are expected to occur about every 500bp of random sequence (1/ (3*%55.2*%55.2)). We note that the theoretical targeting range in random DNA sequence for ZFNs made using the proprietary Sangamo platform has been reported to be 1 in ~31 bp of random sequence, [13] a capability currently superior to that of CoDA. Therefore, an important priority for future experiments will be to identify additional context-sensitive finger units for CoDA to further expand the range of potentially targetable genes and sequences.

Modified ZiFiT software for identifying potential CoDA ZFN target sites

Our new ZiFiT V3.3 program can be accessed at: http://bindr.gdcb.iastate.edu/ZiFiT/ or at http://www.zincfingers.org/software-tools.htm. Output from the ZiFiT program (Supplementary Figure 5) provides the sequence of potential CoDA target sites and unique identification numbers for requesting plasmids encoding the finger units required to assemble

arrays (individual CoDA zinc finger units can be requested from the Joung lab by any interested academic researcher). Alternatively, ZiFiT can also generate DNA sequences encoding CoDA zinc finger arrays required to target a given site; these <290 bp DNA fragments can be synthesized through a commercial provider and then seamlessly cloned into existing Zinc Finger Consortium ZFN expression vectors. [1, 7, 14, 15]

Supplementary References

1. Maeder, M.L., Thibodeau-Beganny, S., Sander, J.D., Voytas, D.F. & Joung, J.K. Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays. Nat Protoc 4, 1471-1501 (2009).

2. Wright, D.A. et al. Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly. Nat Protoc 1, 1637-1652 (2006).

3. Mandell, J.G. & Barbas, C.F., 3rd Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. Nucleic Acids Res 34, W516-523 (2006).

4. Liu, Q., Xia, Z., Zhong, X. & Case, C.C. Validated zinc finger protein designs for all 16 GNN DNA triplet targets. J Biol Chem 277, 3850-3856 (2002).

5. Bae, K.H. et al. Human zinc fingers as building blocks in the construction of artificial transcription factors. Nat Biotechnol 21, 275-280 (2003).

6. Ramirez, C.L. et al. Unexpected failure rates for modular assembly of engineered zinc fingers. Nat Methods 5, 374-375 (2008).

7. Foley, J.E. et al. Rapid Mutation of Endogenous Zebrafish Genes Using Zinc Finger Nucleases Made by Oligomerized Pool ENgineering. PLoS ONE 4, e4348 (2009).

8. Meng, X., Noyes, M.B., Zhu, L.J., Lawson, N.D. & Wolfe, S.A. Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases. Nat Biotechnol 26, 695-701 (2008).

9. Doyon, Y. et al. Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases. Nat Biotechnol 26, 702-708 (2008).

10. Zhang, F. et al. High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. Proc Natl Acad Sci U S A 107, 12028-12033 (2010).

11. Smith, J. et al. Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. Nucleic Acids Res 28, 3361-3369 (2000).

12. Handel, E.M., Alwin, S. & Cathomen, T. Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity. Mol Ther 17, 104-111 (2009).

13. Shukla, V.K. et al. Precise genome modification in the crop species Zea mays using zinc finger nucleases. Nature 459, 337-338 (2009).

14. Maeder, M.L. et al. Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. Mol Cell 31, 294-301 (2008).

15. Foley, J.E. et al. Targeted mutagenesis in zebrafish using customized zinc-finger nucleases. Nat Protoc 4, 1855-1867 (2009).

APPENDIX B. ZIFIT (ZINC FINGER TARGETER): AN UPDATED ZINC FINGER ENGINEERING TOOL

Jeffry D. Sander, Morgan L. Maeder, Deepak Reyon, Daniel F. Voytas, J. Keith Joung & Drena Dobbs. Nucleic Acids Research. 2010 April 30.

ABSTRACT

ZiFiT (Zinc Finger Targeter) is a simple and intuitive web-based tool that provides an interface to identify potential binding sites for engineered zinc finger proteins (ZFPs) in user-supplied DNA sequences. In this updated version, ZiFiT identifies potential sites for ZFPs made by both the modular assembly and OPEN engineering methods. In addition, ZiFiT now integrates additional tools and resources including scoring schemes for modular assembly, an interface with the Zinc Finger Database (ZiFDB) of engineered ZFPs, and direct querying of NCBI BLAST servers for identifying potential off-target sites within a host genome. Taken together, these features facilitate design of ZFPs using reagents made available to the academic research community by the Zinc Finger Consortium. ZiFiT is freely available on the web without registration at http://bindr.gdcb.iastate.edu/ZiFiT/.

INTRODUCTION

Engineered zinc finger proteins (ZFPs) are important tools for gene regulation and genome modification because they can be used to target functional domains to virtually any desired location in a complex genome[1, 2]. Zinc Finger Nucleases (ZFNs) consist of an engineered ZFP fused to a non-specific nuclease domain and can be used to create double-stranded breaks (DSBs) in specific endogenous genes [3]. These DSBs can be exploited to induce highly efficient insertion or alteration of DNA sequences via homologous recombination at the targeted locus [4-11]. Alternatively, imperfect repair of a ZFN-induced DSB by non-homologous end joining can lead to highly efficient generation of gene-specific knockouts [7, 12-17].

Engineered C2H2 zinc finger proteins comprise multiple (usually 3 to 6) ZF domains joined together by a fixed amino acid linker sequence(s), typically TGEKP. Each individual

domain conforms to the zinc finger motif X_2-C-X_{2-4}-C-X_{12}-H_{3-5}-H which, when chelated with a zinc ion, forms a ββα fold. This structure presents a stabilized α-helix (the recognition helix) capable of making base specific contacts with approximately 3 bases in the major groove of double-stranded DNA. Adjacent ZF domains in a ZFP typically specify adjacent DNA triplets, establishing specificity for an extended target site [18].

ZFPs can be engineered to recognize new DNA sequences by altering as many as six important residues in the recognition helix. Because ZFP libraries covering this sequence space (20^{18} variants for a three finger protein) cannot be built or adequately sampled using existing molecular biology techniques, much effort has been placed on developing alternative methods for engineering multi-finger proteins [7, 19-29]. These methods typically involve identifying individual zinc finger domains that recognize specific DNA triplet "subsites" and then joining these domains together to create multi-finger proteins. The two most common ZFP engineering methods, modular assembly and oligomerized pool engineering (OPEN), both use variations on this general approach (Reviewed in Cathomen and Joung, 2008, Molecular Therapy). Modular assembly usually assumes that a single domain (module) can recognize a specific DNA triplet regardless of the position of the triplet within the target site or the identities of adjacent neighboring fingers (i.e., it assumes binding of the ZF module is context-independent). Appropriate modules are simply joined together to create a ZFP that should recognize the target sequence. Modular assembly is relatively simple to accomplish; however, ZFPs generated using this method have been shown to have a high failure rate *in vivo* [30-32]. In contrast, OPEN uses customized "pools" of ZF modules selected to recognize triplets in a specific sequence context. These pools can be assembled to create combinatorial libraries (with up to one million unique solutions for a three-finger protein) from which the ZFPs best able to bind the chosen target DNA site are identified. Although OPEN is somewhat more labor intensive, it is more robust, with a higher success rate than modular assembly [7].

ZiFiT provides a simple interface for scanning a DNA sequence to identify potential ZFP and ZFN binding sites. The updated version (3.2) identifies target sites for proteins engineered using either OPEN or modular assembly. ZiFiT 3.2 also provides several new tools to help researchers evaluate ZFP targets, including validated scoring schemes for

ranking potential target sites, a tool for querying NCBI BLAST servers for potential off-target sites, and a seamless interface with the Zinc Finger Database (ZiFDB, a database of engineered ZFPs [33]).

MATERIALS & METHODS

Modular Assembly

The modular assembly engineering approach employs individual zinc finger domains (modules) that have been pre-characterized (in the middle position of a three-finger array) to bind a specific DNA triplet sub site. Several (three to six) of these modules can be arranged and linked together to generate a ZFP that recognizes an extended DNA sequence corresponding to the desired target site. ZiFiT provides support for the three most commonly used module sets developed by three independent research groups [21, 23, 27].

Oligomerized Pool ENgineering

Oligomerized Pool ENgineering (OPEN) utilizes pools of zinc finger domains pre-characterized to bind a specific DNA triplet sub site in the first, second or third position of the target site for a three-finger ZFP[7]. Appropriate pools are combined to generate hundreds of thousands of distinct solutions for a given 9 bp target DNA sequence. The best solutions are subsequently identified using a bacterial two-hybrid (B2H) assay [34]. Although this method requires considerable effort, it reliably generates ZFPs that bind with high affinity and specificity to their intended target site [7, 9, 11, 16].

GNN Scoring

The GNN score is an empirical estimate of the probability that a modularly assembled three-finger ZFP will provide > 1.6-fold activation in the B2H assay (proteins that fail to meet this cutoff have been shown to fail to function in mammalian cells [30]). Using probabilities based on evaluation of a set of 168 three-finger ZFPs generated by modular assembly, ZFPs designed to bind target sites containing 3, 2, 1, and 0 triplets of the form GNN (where N is any nucleotide) are predicted to have success rates of 59%, 29%, 12%, and 0%, respectively[30].

Affinity Scoring

The affinity score is an energy-based parameter that predicts which modularly assembled three-finger ZFPs are most likely to function by inferring the contributions of the individual modules. Scores are calculated by estimating the relative free energy contributions of individual modules from dissociation constants reported for modules in the middle (F2) position of a three-finger ZFP[27]. These affinity scores have been calibrated to B2H fold-activation values for modularly assembled ZFPs tested *in vivo*. ZFPs with affinity scores less than 5 are expected to have adequate affinity to function in the B2H assay [35]. These scores do not directly address specificity and are available only for ZFPs composed exclusively of Barbas GNN and TGG modules.

Program Input

The ZiFiT 3.2 interface enables customizable searches for potential ZFP and ZFN binding sites that can be targeted using either the modular assembly or OPEN engineering methods. After selecting *ZiFiT* from the menu bar, users select their preferred engineering method (modular assembly or OPEN) and target type (ZFP or ZFN), e.g., OPEN - Zinc Finger Nuclease. In all interfaces, users enter their DNA query sequence into the *Sequence* input box near the top of the page (Figure B.1). Sequences can be submitted either in FASTA format or raw text. Ideally, sequences should be entered using uppercase characters to denote exons and lowercase characters to denote introns. Entering information in this format can facilitate target selection for certain experimental applications, such as generation of knockout mutations via ZFN-induced DSBs within a protein-encoding region. This function can be disabled by de-selecting the *Exon/Intron Case Sensitivity* check box immediately

below the *Sequence* input box.

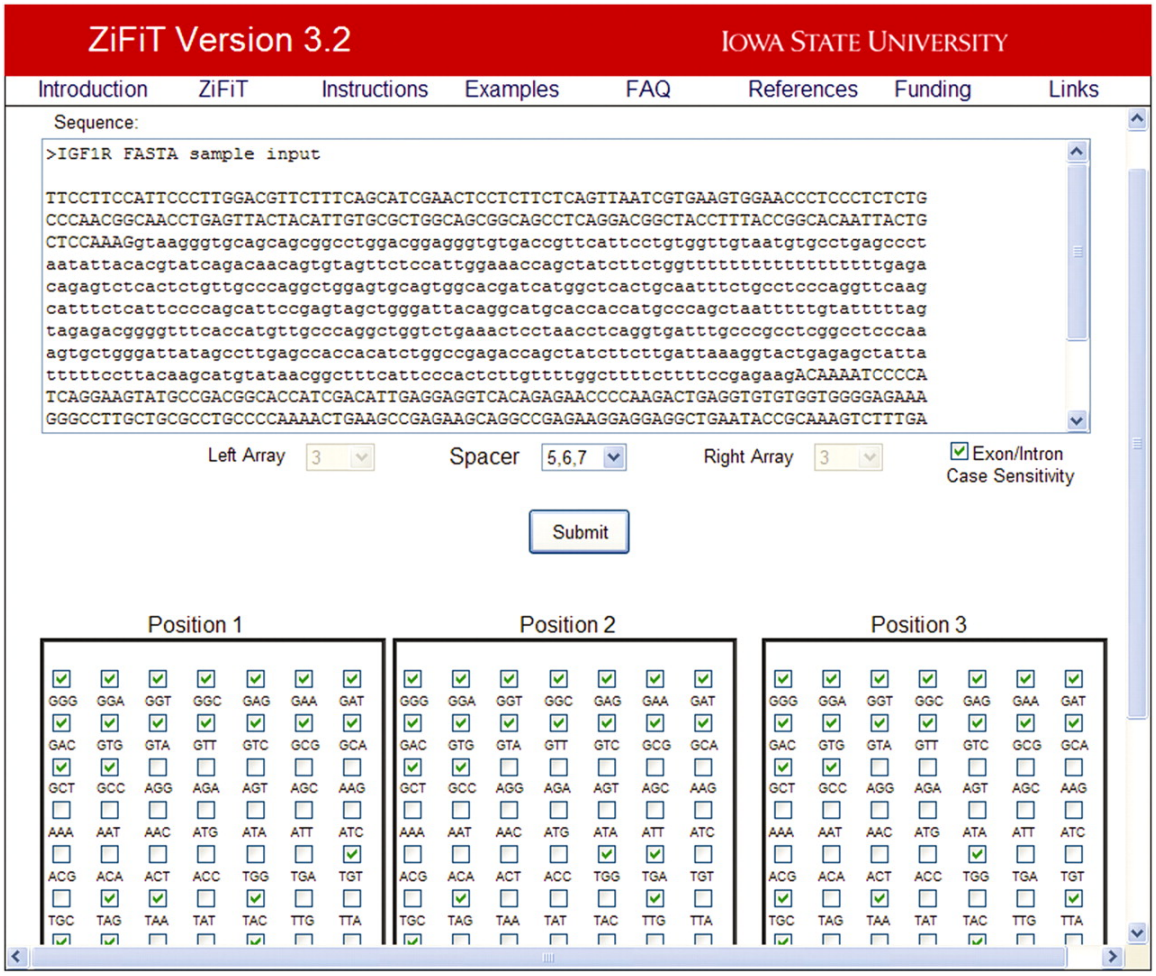


Figure B.1. ZiFiT Input Window. In this example the user has chosen to scan for ZFN sites using OPEN. The sequence has been entered in FASTA format, with exon sequences in uppercase and intron sequences in lowercase. Published sets of OPEN finger pools available through the Zinc Finger Consortium are checked by default.

The engineering method chosen by the user determines which sequences can be targeted. For modular assembly, users can choose one or more of the three available module sets by selecting the *Barbas*, *ToolGen*, and *Sangamo* check boxes at the top of the page[21, 23, 26, 27, 36]. Several studies have generated functional zinc fingers by combining modules from different sets [21]. For OPEN, users indicate which pools they wish to use by choosing the corresponding target DNA triplet for each finger position in a three-finger array. All published OPEN pools currently available from the Zinc Finger Consortium are checked by default [7].

Users can specify the number of ZF modules to include in the *Left Array* and *Right Array* using drop-down menus below the *Sequence* input box. ZiFiT restricts the number of modules in concordance with the available reagents. OPEN reagents and Sangamo modules are specific to three-finger ZFPs or ZFNs, whereas practitioners of modular assembly implementing only Barbas and Toolgen modules may scan for individual target sites for ZFPs consisting of 3 to 8 fingers, or for dimeric target sites for ZFNs consisting of 3 or 4 fingers for each target "half-site." Users should note that although increasing the number of modules might be expected to confer enhanced specificity, this is not always the case because deformation of the DNA upon ZFP binding may limit the number of ZF domains that bind concurrently [37]. In addition, because three domains are often sufficient to bind DNA with high affinity, longer ZFP arrays may require disrupted linkers between domains to prevent binding at unintended sub-sequences within the target site [24].

For ZFN targets, users must also specify the length of the spacer region between the binding sites for the left and right ZF arrays using the *Spacer* drop-down menu immediately below the *Sequence* input box. An active dimeric ZFN cleaves within the spacer region and its preferred length is dependent on the sequence and length of the amino acid linker between the ZF and nuclease domains. Zinc Finger Consortium vectors harbor a linker that works with spacers of 5 or 6 bp of DNA [38-40]. An additional linker permitting spacers with 6 or 7 bp of DNA has also been identified[40]. By default, ZiFiT scans for ZFN targets with spacers of 5, 6, or 7 bp. Users can choose to limit the scan to only one or two spacer lengths.

Advanced search options are accessible by selecting the *Advanced* link in the lower right hand corner (this link then toggles to *Basic*, which hides the Advanced options). Advanced options allow users to customize their scan by adjusting additional constraints. For example, users can restrict the minimum and maximum number of GNN, ANN, CNN and TNN triplets for reported targets. This feature can help users identify the best sites by restricting searches to more successful GNN-rich sites [12, 30]. Additional Advanced options available for a subset of the input interfaces include: (i) *Ignore Asp overlap:* This option is available to users of modular assembly; it refers to "target site overlap" in which an Asp in position +2 of the recognition helix specifies a 4^{th} base[41]. This option is useful for troubleshooting why ZiFiT fails to return an expected target site; its use during design should

be restricted to advanced users. (ii) *Search both strands:* Because ZFPs bind directionally in a 5’ - 3’ manner, they can be engineered to bind either strand of DNA. ZiFiT searches both strands by default. Additional guidance for using Advanced options is provided on the ZiFiT Instructions and FAQ pages.

Program Output

ZiFiT scans the user-supplied input DNA sequence for potential ZFP binding or ZFN cleavage sites based on the selected engineering method and any additional user-defined restrictions. For each user submission, ZiFiT displays a graphic map of the submitted sequence with each target site (‘hit’) indicated above the sequence (Figure B.2a). (Users may need to “enable pop-ups” in their browser for this feature to function properly.) The submitted sequence is displayed as a red bar at the bottom of the map. When an *Exon/Intron Case-Sensitive* search is performed (see Program Input), exons are represented by thick red bars and introns by thin red lines. Hits are represented as short colored bars above the sequence track, with overlapping hits overflowing vertically into auxiliary tracks. Each bar is a clickable link to detailed target information on the main output page. For ZFN scans, hits are color-coded according to the length of the spacer.

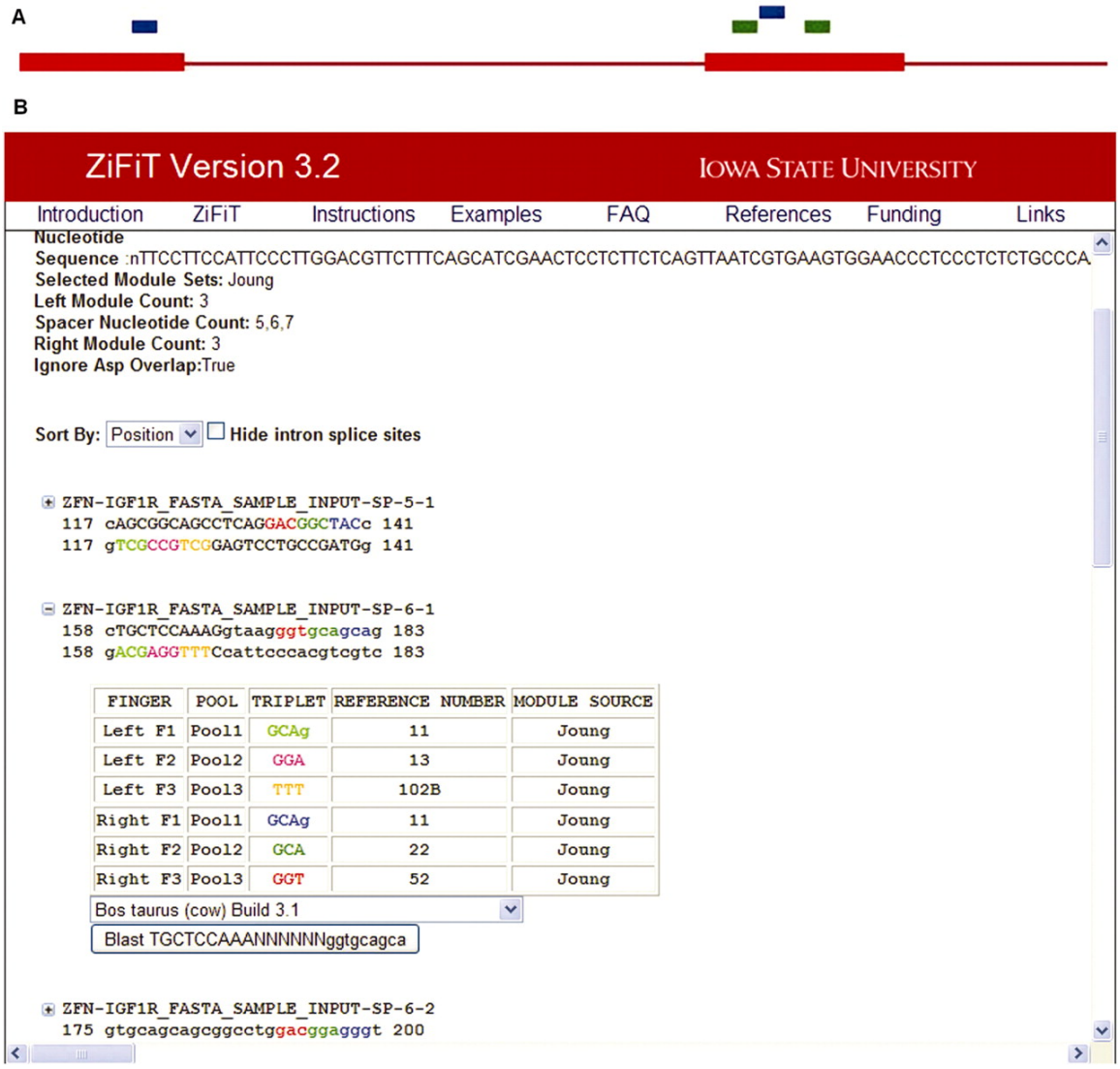


Figure B.2. ZiFiT Output Windows. (**A**) The Graphic Summary window (top) displays potential ZFN-binding sites ('hits') identified in the input sequence. Exons (denoted by uppercase characters in the input sequence) are displayed as thick red bars and introns as thin red lines. Hits are represented by short colored bars above the input sequence track; these serve as bookmarks that are linked to individual target sites in the Detailed Target List. ZFN hits are color-coded based on spacer size (5 bp = Blue; 6 bp = Green; 7 bp = Gold). (**B**) The Detailed Target List window (bottom) provides in-depth information about each hit. Hits are presented as double-stranded DNA sequences and labeled according to the FASTA description, spacer size, and index. Each hit can be expanded to reveal a reagent list for generating the corresponding ZFP. Hits can be sorted according to various criteria as detailed in the text.

The main output page opens with a summary of the search parameters, followed by a drop-down *Sort By* menu that can be used to sort individual hits based on position or score

when available (see Materials & Methods). Because ZFN targets consist of two ZFP target sites, both of which must be targeted successfully, score-based sorting considers the score of the inferior scoring array site before the better scoring array target. When users implement an 'Exon/intron case-sensitive' scan for ZFN targets, a *Filter intronic splice sites* checkbox is present immediately to the right of the *Sort By* menu. Selecting this box will hide ZFN targets whose spacers occur within (or overlap) an intron (Figure 2b). In addition to its web-based output, ZiFiT also provides a text version of the output which can be downloaded as a .csv file from the top of the output page.

Each hit is named using the description/comment line of the submitted FASTA sequence and an index number. If no sequence name is supplied by the user, this parameter is set to 'Unknown'. Names for ZFN targets also include the spacer length. For example, for a submission with the FASTA description '>ZFN-SAMPLE," the third ZFN target site with a spacer of 7 bps would be labeled 'ZFN-SAMPLE-SP-7-3'. Immediately beneath each hit name, the double-stranded DNA target sequence is displayed, along with its position within the submitted sequence. Individual triplets within this sequence are highlighted with distinct colors denoting the targets of individual ZF domains. Each highlighted ZFP is linked to ZiFDB, a database of engineered ZFPs. Clicking these links automatically queries ZiFDB for available information regarding ZFPs tested against the same or similar target DNA sequences [33]. If an expected functional activity score is available for a given ZFP, the score is presented on the same line as the target. ZiFiT currently provides two validated functional scoring schemes for modular assembly targets [30, 35]. Scoring schemes for OPEN targets are under development.

Individual ZF targets can be expanded using the '+' to the left of the target name. Expanding a target reveals three types of information. (i) A table describing reagents that can be used to generate corresponding ZFPs. Each row in the table describes a reagent (pool or module) corresponding to a *Triplet* DNA sequence, which is color-coded according to its position in the double-stranded sequence above the table. Entries in the *Reference Number* column of the table signify the names of reagents (either modules or pools) that are currently available to the academic research community through the Zinc Finger Consortium. Modules (and other reagents for performing modular assembly) are available from the non-

profit plasmid distribution service Addgene (http://www.addgene.org/zfc) [42]. OPEN pools are available by request from the Joung lab (jjoung@partners.org) and other reagents for practicing OPEN are also available from Addgene (http://www.addgene.org/zfc). (ii) Sequences of oligonucleotides that must be synthesized to create bacterial two-hybrid selection and/or reporter strains needed to screen or select ZFPs for DNA-binding activity [34, 42]. (iii) An *Organism* drop-down menu for selecting a host genome and BLAST button that can be used to scan the selected host organism genome for exact & similar target matches. The BLAST button submits search parameters to NCBI and via a popup directs users to NCBI website where they can initiate the query by selecting the "view report" button. This is useful because it is generally desirable to avoid targeting sites that occur frequently in a genome (e.g., sites that fall within repeat regions). When using BLAST to search for genomic ZFN targets, the spacer is replaced with N's to prevent it from positively influencing a scan. Due to the nature of the algorithm (and a fixed spacer size in the case of a nuclease), this query is not guaranteed to identify all similar sites. ZiFiT output may need to complete loading before BLAST queries are accessible. ZiFiT is freely available on the web without registration at http://bindr.gdcb.iastate.edu/ZiFiT/.

ACKNOWLEDGEMENTS

We respectfully acknowledge those who have shared their results with the zinc finger research community. J.D.S. is supported by the National Institutes of Health [T32CA009216]. M.L.M. is supported by a National Science Foundation Graduate Research Fellowship [2009080622]. D.F.V is supported by the National Science Foundation [DBI 0923827, DBI 0501678]. J.K.J. is supported by the National Institutes of Health [R01 GM069906, R01 GM072621, R01 GM088040], the NSF [DBI 0923827], and the Massachusetts General Hospital Pathology Service. D.D. and D.R. are supported by the National Science Foundation [DBI 0923827].

Conflict of interest statement. None declared.

REFERENCES

1. Carroll, D., *Progress and prospects: zinc-finger nucleases as gene therapy agents.* Gene Ther, 2008. **15**(22): p. 1463-8.
2. Cathomen, T. and J. Keith Joung, *Zinc-finger nucleases: the next generation emerges.* Mol Ther, 2008. **16**(7): p. 1200-7.
3. Kim, Y.G., J. Cha, and S. Chandrasegaran, *Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain.* Proc Natl Acad Sci U S A, 1996. **93**(3): p. 1156-60.
4. Bibikova, M., et al., *Enhancing gene targeting with designed zinc finger nucleases.* Science, 2003. **300**(5620): p. 764.
5. Hockemeyer, D., et al., *Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases.* Nat Biotechnol, 2009. **27**(9): p. 851-7.
6. Lombardo, A., et al., *Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery.* Nat Biotechnol, 2007. **25**(11): p. 1298-306.
7. Maeder, M.L., et al., *Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification.* Mol Cell, 2008. **31**(2): p. 294-301.
8. Shukla, V.K., et al., *Precise genome modification in the crop species Zea mays using zinc-finger nucleases.* Nature, 2009. **459**(7245): p. 437-41.
9. Townsend, J.A., et al., *High-frequency modification of plant genes using engineered zinc-finger nucleases.* Nature, 2009. **459**(7245): p. 442-5.
10. Urnov, F.D., et al., *Highly efficient endogenous human gene correction using designed zinc-finger nucleases.* Nature, 2005. **435**(7042): p. 646-51.
11. Zou, J., et al., *Gene targeting of a disease-related gene in human induced pluripotent stem and embryonic stem cells.* Cell Stem Cell, 2009. **5**(1): p. 97-110.
12. Bibikova, M., et al., *Targeted chromosomal cleavage and mutagenesis in Drosophila using zinc-finger nucleases.* Genetics, 2002. **161**(3): p. 1169-75.
13. Perez, E.E., et al., *Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(7): p. 808-16.
14. Meng, X., et al., *Targeted gene inactivation in zebrafish using engineered zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(6): p. 695-701.
15. Doyon, Y., et al., *Heritable targeted gene disruption in zebrafish using designed zinc-finger nucleases.* Nat Biotechnol, 2008. **26**(6): p. 702-8.
16. Foley, J.E., et al., *Rapid mutation of endogenous zebrafish genes using zinc finger nucleases made by Oligomerized Pool ENgineering (OPEN).* PLoS ONE, 2009. **4**(2): p. e4348.
17. Lee, H.J., E. Kim, and J.S. Kim, *Targeted chromosomal deletions in human cells using zinc finger nucleases.* Genome Res, 2010. **20**(1): p. 81-9.
18. Pavletich, N.P. and C.O. Pabo, *Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A.* Science, 1991. **252**(5007): p. 809-17.
19. Carroll, D., et al., *Design, construction and in vitro testing of zinc finger nucleases.* Nat Protoc, 2006. **1**(3): p. 1329-41.
20. Hurt, J.A., et al., *Highly specific zinc finger proteins obtained by directed domain shuffling and cell-based selection.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 12271-6.

21. Bae, K.H., et al., *Human zinc fingers as building blocks in the construction of artificial transcription factors.* Nat Biotechnol, 2003. **21**(3): p. 275-80.
22. Sera, T. and C. Uranga, *Rational design of artificial zinc-finger proteins using a nondegenerate recognition code table.* Biochemistry, 2002. **41**(22): p. 7074-81.
23. Liu, Q., et al., *Validated zinc finger protein designs for all 16 GNN DNA triplet targets.* J Biol Chem, 2002. **277**(6): p. 3850-6.
24. Moore, M., A. Klug, and Y. Choo, *Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units.* Proc Natl Acad Sci U S A, 2001. **98**(4): p. 1437-41.
25. Isalan, M., A. Klug, and Y. Choo, *A rapid, generally applicable method to engineer zinc fingers illustrated by targeting the HIV-1 promoter.* Nat Biotechnol, 2001. **19**(7): p. 656-60.
26. Dreier, B., et al., *Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors.* J Biol Chem, 2001. **276**(31): p. 29466-78.
27. Segal, D.J., et al., *Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences.* Proc Natl Acad Sci U S A, 1999. **96**(6): p. 2758-63.
28. Greisman, H.A. and C.O. Pabo, *A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites.* Science, 1997. **275**(5300): p. 657-61.
29. Wu, H., W.P. Yang, and C.F. Barbas, 3rd, *Building zinc fingers by selection: toward a therapeutic application.* Proc Natl Acad Sci U S A, 1995. **92**(2): p. 344-8.
30. Ramirez, C.L., et al., *Unexpected failure rates for modular assembly of engineered zinc fingers.* Nat Methods, 2008. **5**(5): p. 374-5.
31. Kim, H.J., et al., *Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly.* Genome Res, 2009. **19**(7): p. 1279-88.
32. Joung, J.K., D.F. Voytas, and T. Cathomen, *Reply to "Genome editing with modularly assembled zinc-finger nucleases".* Nat Methods, 2010. **7**(2): p. 91-2.
33. Fu, F., et al., *Zinc Finger Database (ZiFDB): a repository for information on C2H2 zinc fingers and engineered zinc-finger arrays.* Nucleic Acids Res, 2009. **37**(Database issue): p. D279-83.
34. Maeder, M.L., et al., *Oligomerized pool engineering (OPEN): an 'open-source' protocol for making customized zinc-finger arrays.* Nat Protoc, 2009. **4**(10): p. 1471-501.
35. Sander, J.D., et al., *An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins.* Nucleic Acids Res, 2009. **37**(2): p. 506-15.
36. Dreier, B., et al., *Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors.* J Biol Chem, 2005. **280**(42): p. 35588-97.
37. Peisach, E. and C.O. Pabo, *Constraints for zinc finger linker design as inferred from X-ray crystal structure of tandem Zif268-DNA complexes.* J Mol Biol, 2003. **330**(1): p. 1-7.

38. Smith, J., et al., *Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains.* Nucleic Acids Res, 2000. **28**(17): p. 3361-9.
39. Bibikova, M., et al., *Stimulation of homologous recombination through targeted cleavage by chimeric nucleases.* Mol Cell Biol, 2001. **21**(1): p. 289-97.
40. Handel, E.M., S. Alwin, and T. Cathomen, *Expanding or restricting the target site repertoire of zinc-finger nucleases: the inter-domain linker as a major determinant of target site selectivity.* Mol Ther, 2009. **17**(1): p. 104-11.
41. Elrod-Erickson, M., T.E. Benson, and C.O. Pabo, *High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition.* Structure, 1998. **6**(4): p. 451-64.
42. Wright, D.A., et al., *Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly.* Nat Protoc, 2006. **1**(3): p. 1637-52.

APPENDIX C. SUPPLEMENTARY MATERIALS

Construction of TALE repeat arrays and TALE nuclease expression vectors

We assembled DNA encoding TALE repeat arrays that possess the same sequence architecture described by Miller et al. [1] in which four TALE repeat backbones that differ in their DNA and amino acid sequences (and which we designate types I, II, III, and IV; see Supplementary Figure 6) occur in a consistent repeating pattern (e.g.—I-II-III-IV-III-III-IV-I-II-III-IV). To do this, we initially commercially synthesized (Genscript) a large series of plasmids encoding the various individual TALE repeats required to construct the arrays shown in Figure 1 and Supplementary Figures 3, 4, and 5). These individual TALE repeats harbored the following repeat variable di-residues (RVDs): NN (for binding to G), NG (for binding to T), NI (for binding to A), or HD (for binding to C) (Supplementary Table 1 and Supplementary Figure 7). In all of these plasmids, the DNA sequence encoding the individual TALE repeat is flanked by unique XbaI and BbsI restriction sites on the 5' end and unique BsaI and BamHI restriction sites on the 3' end (Supplementary Table 1). In addition, for all of these plasmids, the 4 bp overhang generated upon digestion with BsaI can be ligated to the 4 bp overhang generated upon digestion with BbsI (Supplementary Table 1). Note that in order to enable cloning of the final DNA fragment encoding the full-length TALE repeat array, slightly different 5' and 3' ends were required for Type I TALE repeats present on the N-terminal end of an array and for Type III TALE repeats on the C-terminal end of an array, respectively (Supplementary Table 1). The full DNA and amino acid sequences of these commercially synthesized plasmids are given in Supplementary Table 1 and Supplementary Figure 7.

To join together sequences encoding individual TALE repeats, we used the restriction enzyme-based strategy outlined in Supplementary Figure 8, an approach similar to one we and others have used previously to fuse individual zinc fingers into arrays [2, 3]. Briefly, DNA fragments encoding first and second TALE repeats can be joined together by ligating compatible overhangs generated by digestion of the first TALE repeat-encoding DNA with BsaI and by digestion of the second TALE repeat-encoding DNA with BbsI (Supplementary Figure 8). In the resulting fusion, the TALE repeats are joined together in the order first-

then-second (amino-terminal to carboxy-terminal). In addition, the DNA encoding two TALE repeats remains flanked by a BbsI site on the 5' end and a BsaI site on the 3' end (Supplementary Figure 8). Thus, this process can continue to be repeated in an iterative fashion to join together both single and multiple TALE repeats into longer arrays. To generate the variable length TALE repeat arrays required for our studies, we joined together subsets of TALE repeats into shorter arrays and then joined these together to create the final desired full-length arrays. An example of how we utilized this strategy is illustrated for creation of the TALE repeat array for TALE nuclease clone #1257 (targeted to the hey2 gene) in Supplementary Figure 9.

We also constructed TALE nuclease expression vectors into which our assembled TALE repeat arrays could be cloned. These vectors each harbor the following elements: a T7 promoter, a nuclear localization signal, a FLAG tag, amino acids 153 to 288 from the TALE13 protein (numbering as defined by Miller et al. [1]), two adjacent BsmBI restriction sites into which a DNA fragment encoding a TALE repeat array can be cloned, a 0.5 TALE repeat, amino acids 715 to 777 from the C-terminal end of the TALE13 protein (numbering as defined by Miller et al [1]), and the wild-type FokI cleavage domain. Plasmid pJDS70 encodes a 0.5 TALE repeat with a NI RVD (for recognition of an A nucleotide), plasmid pJDS71 encodes a 0.5 TALE repeat with a HD RVD (for recognition of a C nucleotide), and plasmid pJDS78 encodes a 0.5 TALE repeat with a NG RVD (for recognition of a T nucleotide). The full DNA sequences of these plasmids are provided in Supplementary Figure 10. BbsI/BsaI-digested DNA fragments encoding arrays of TALE repeats were cloned into BsmBI-digested TALE nuclease expression vector backbone (the overhangs generated by digestion of the TALE repeat array-encoding fragment with BbsI and BsaI are complementary to the first and second BsmBI overhangs, respectively, generated by digestion of the TALE nuclease expression vector). In the resulting plasmids, the TALE repeat arrays are expressed from the T7 promoter on RNA as in-frame fusions with the NLS, the FLAG tag, and the wild-type FokI nuclease domain. The full DNA and amino acid sequences of the TALE nucleases used in this study can be found in Supplementary Figures 4 and 5. All of the plasmids described above are available upon request from the Joung lab (http://www.jounglab.org).

Preparation of ZFN- and TALE nuclease-encoding mRNAs

ZFN and TALE nuclease expression vectors were linearized with PmeI and transcribed in vitro using the mMESSAGE mMACHINE£ T7 ULTRA kit (Ambion). The transcribed RNAs were then polyadenylated using the reagents in the same kit. Microinjection and analysis of somatic mutations Approximately 2 nl of the TALE nuclease mRNAs (300 pg/nl) or ZFN mRNAs (200-250 pg/nl) were injected into one-cell stage zebrafish embryos. On the next day, the surviving injected embryos were grouped into either “normal” or “deformed” phenotypes. Genomic DNA was extracted from pools of 8-10 embryos from each “normal” group. Target genomic loci were amplified using primers designed to anneal approximately 150 to 200 base pairs upstream and downstream from the expected cut site. The resulting PCR product was cloned into pGEM-5Zf(+) using a pGEM-T Easy kit (Promega) or into pCR4 TOPO-TA using a TOPO-TA kit (Invitrogen). After transformation of the ligated vectors, single colonies were reinoculated for plasmid DNA isolation and sequencing. The sequence of each clone represents one amplified allele.

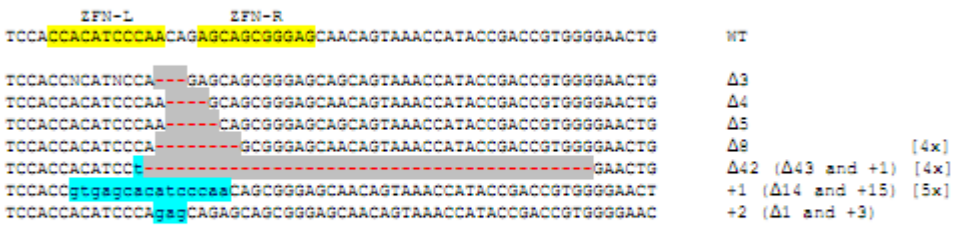
Supplementary References

1. Miller, J.C., et al., *A TALE nuclease architecture for efficient genome editing.* Nat Biotechnol, 2011. **29**(2): p. 143-8.
2. Bae, K.H., et al., *Human zinc fingers as building blocks in the construction of artificial transcription factors.* Nat Biotechnol, 2003. **21**(3): p. 275-80.
3. Wright, D.A., et al., *Standardized reagents and protocols for engineering zinc finger nucleases by modular assembly.* Nat Protoc, 2006. **1**(3): p. 1637-52.

Supplementary Figure 1

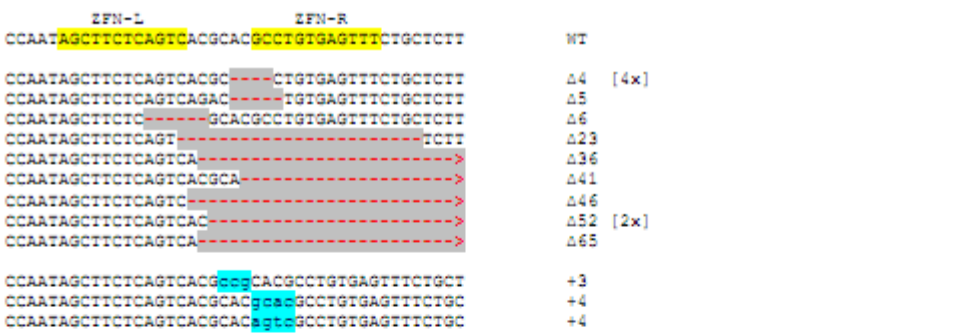
hey2:

MLM407/169-2 Mutations in 17 of 59 sequences: ~29%



gria3a:

FFL47/FFL94 Mutations in 16 of 62 sequences: ~26%



Endogenous zebrafish gene mutations introduced by zinc finger nucleases (ZFNs)

Target sequences, frequencies of mutations, and sequences of mutations induced by ZFNs in embryonic zebrafish cells. For each pair of ZFNs, the wild-type (WT) target sequence is shown at the top with ZFN-binding sites highlighted in yellow. Deletions are indicated by gray highlighted red dashes and insertions by blue highlighted lower case blue letters. The sizes of the insertions (+) or deletions (Δ) are indicated to the right of each mutant allele. The number of times that each mutant allele was isolated is shown in brackets. Mutation frequencies are calculated as the number of mutant alleles isolated/the total number of alleles analyzed. Full DNA and amino acid sequences of the ZFNs used to induce these mutations are shown in **Supplementary Figure 2**.

Supplementary Figure 2

A.

For *hey2*:

MLM407:

TCTAGACCCGGGGAGCGCCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGAGGAACTTCATCCTTCAGAGGCA
TACCCGTACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCTGCTGCACAACTTGAC
GCGGCATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCAGTCGGAGCGAC
CACCTGAGCCTGCACCTAAAAACCCACCTGAGGGGATCC

169-2:

TCTAGACCCGGGGAGCGCCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGGACAACGCCCACCTTGCGCGCC
ATACCCGTACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCGCGAGGACACGTTG
ACCCGCCATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCAGTCAGAACG
GGACGCTGACCCGGCACCTAAAAACCCACCTGAGGGGATCC

For *gria3a*:

FFL47:

TCTAGACCCGGGGAGCGACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCAGCCCGAACGGGTTGCAGGTGC
ATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCAGTAGCGCGAGCAACCTG
ACCCGCCACCTAAAAACCCACACCGGCTCTCAGAAGCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGACCAA
GCAGAAGCTTACCCTGCATACCCGTACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTC
CGACAAGAGCGTGTTGCGGCGGCATCTACGTACGCACCTGAGGGGATCC

FFL94:

TCTAGACCCGGGGAGCGCCCCTTCCAGTGTCGCATTTGCATGCGGAACTTTTCGAGCGCCCAGGCCCTTGCCCGGC
ATACCCGTACTCATACCGGTGAAAAACCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCAGGGCGGCAACTTG
ACCAGGCATCTACGTACGCACACCGGCTCTCAGAAGCCGTTTCAGTGTCGGATCTGTATGCGAAATTTCTCCCAGCG
GGAGCACTTGGAGGTCCATCTACGTACGCACACCGGCGAGAAGCCATTCCAATGCCGAATATGCATGCGCAACTTCA
GTGACCCCAGCAACCTGCGGCGCCACCTAAAAACCCACCTGAGGGGATCC

B.

For *hey2*:

MLM407:

SRPGERPFQCRICMRNFSRNFILQRHTRTHTGEKPFQCRICMRNFSLLHNLTRHLRTHTGEKPFQCRICMRNFSRSDHLS
LHLKTHLRGS

169-2:

SRPGERPFQCRICMRNFSDNAHLARHTRTHTGEKPFQCRICMRNFSREDTLTRHLRTHTGEKPFQCRICMRNFSQNGTL
TRHLKTHLRGS

For *gria3a*:

FFL47:

SRPGERPFQCRICMRNFSSPNGLQVHLRTHTGEKPFQCRICMRNFSSASNLTRHLKTHTGSQKPFQCRICMRNFSTKQKL
TLHTRTHTGEKPFQCRICMRNFSDKSVLRRHLRTHLRGS

FFL94:

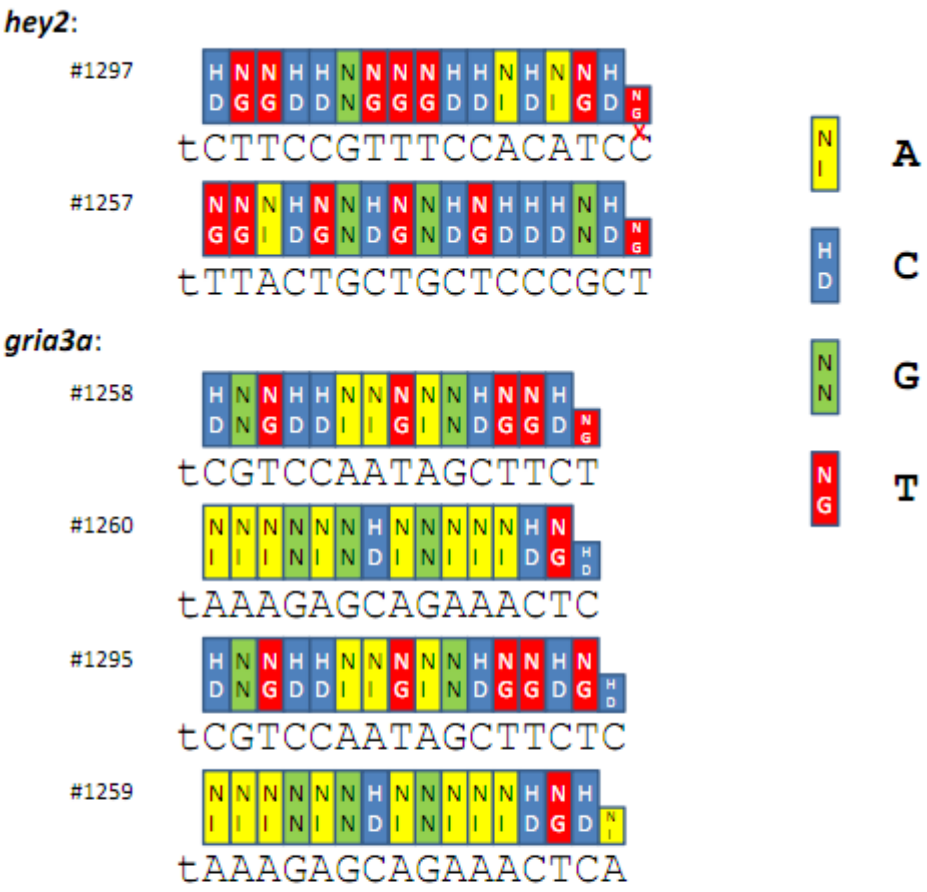
SRPGERPFQCRICMRNFSSAQALARHTRTHTGEKPFQCRICMRNFSQGGNLTRHLRTHTGSQKPFQCRICMRNFSQREH
LEVHLRTHTGEKPFQCRICMRNFSDPSNLRRHLKTHLRGS

DNA and amino acid sequences of ZFNs used in this study

A. DNA sequences encoding the ZFNs used in this study. Naming of ZFNs is as described in **Supplementary Figure 1**. The *Xba*I-*BamH*I fragments shown above were cloned into expression vectors pMLM290 and pMLM292 (full sequences of these plasmids are available online at http://www.addgene.org/zfc, click on the "Nuclease Expression Vectors" link).

B. Amino acid sequences of the zinc finger arrays used in this study. The zinc finger arrays shown were used to make the ZFNs targeted to *hey2* and *gria3a*. These amino acid sequences correspond to the DNA sequences given above in A.

Supplementary Figure 3



Schematic overview of TALE repeat arrays and cognate binding sites engineered for this study

Individual TALE repeat arrays are shown as colored rectangles with the RVDs designated by two letters. C-terminal 0.5 repeats are shown as smaller colored rectangles. The nucleotides bound by each TALE repeat are shown on the right side of the figure. The TALE repeat arrays engineered for each binding site are shown and with their clone names shown to the left (names are as in **Figure 1**). Each DNA site targeted is written 5' to 3' (left to right) with the 5' adjacent T nucleotide shown in lower case. The red "X" shown with TALE repeat array #1297 indicates an inadvertent mismatch between the identity of the 0.5 TALE repeat and the nucleotide at that position. However, consistent with the results of Miller et al. that TALE repeat arrays can bind *in vitro* to targets with one or more mismatches[1], TALE nuclease #1297 can still cleave its target efficiently in zebrafish (**Figure 1**).

Supplementary Figure 4

Common Sequence:

GACGGATCGGGAGATCTCCCGATCCCCTATGGTCGACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCA
GTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACC
GACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATT
GATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTT
ACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGACTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATAT
GCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTT
CCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATA

GCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGAC
TTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAG
CTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAGCTGXXXX
XXCAACTAGTCAAAAGTGAACTGGAGGAGAAGAAATCTGAACTTCGTCATAAATTGAAATATGTGCCTCATGAATATATTG
AATTAATTGAAATTGCCAGAAATTCCACTCAGGATAGAATTCTTGAAATGAAGGTAATGGAATTTTTTATGAAAGTTTATGG
ATATAGAGGTAAACATTTGGGTGGATCAAGGAAACCGGACGGAGCAATTTATACTGTCGGATCTCCTATTGATTACGGTGTG
ATCGTGGATACTAAAGCTTATAGCGGAGGTTATAATCTGCCAATTGGCCAAGCAGATGAAATGCAACGATATGTCGAAGAA
AATCAAACACGAAACAAACATATCAACCCTAATGAATGGTGGAAAGTCTATCCATCTTCTGTAACGGAATTTAAGTTTTTAT
TTGTGAGTGGTCACTTTAAAGGAAACTACAAAGCTCAGCTTACACGATTAAATCATATCACTAATTGTAATGGAGCTGTTCTT
AGTGTAGAAGAGCTTTTAATTGGTGGAGAAATGATTAAAGCCGGCACATTAACCTTAGAGGAAGTGAGACGGAAATTTAAT
AACGGCGAGATAAACTTTTAAGGGCCCTTCGAAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTACCG
GTCATCATCACCATCACCATTGAGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGC
CCCTCCCCCGTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTG
TCTGAGTAGGTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCA
TGCTGGGGATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTGT
AGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTT
TCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGCATCCCTTTAGGGTTCCGAT
TTAGTGCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGT
TTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGG
TCTATTCTTTTGATTTATAAGGGATTTTGGGGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCG
AATTAATTCTGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGGCAGGCAGAAGTATGCAAAGCATGC
ATCTCAATTAGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTA
GTCAGCAACCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCT
GACTAATTTTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGA
GGCCTAGGCTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGATCTGATCAGCACGTGTTGACAATTAATCATCG
GCATAGTATATCGGCATAGTATAATACGACAAGGTGAGGAACTAAACCATGGCCAAGCCTTTGTCTCAAGAAGAATCCACC
CTCATTGAAAGAGCAACGGCTACAATCAACAGCATCCCCATCTCTGAAGACTACAGCGTCGCCAGCGCAGCTCTCTCTAGCG
ACGGCCGCATCTTCACTGGTGTCAATGTATATCATTTTACTGGGGGACCTTGTGCAGAACTCGTGGTGCTGGGCACTGCTGCT
GCTGCGGCAGCTGGCAACCTGACTTGTATCGTCGCGATCGGAAATGAGAACAGGGGCATCTTGAGCCCCTGCGGACGGTGTC
GACAGGTGCTTCTCGATCTGCATCCTGGGATCAAAGCGATAGTGAAGGACAGTGATGGACAGCCGACGGCAGTTGGGATTC
GTGAATTGCTGCCCTCTGGTTATGTGTGGGAGGGCTAAGCACTTCGTGGCCGAGGAGCAGGACTGACACGTGCTACGAGATT
TCGATTCCACCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGG
GGATCTCATGCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAA
ATTTCACAAATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGTATAC
CGTCGACCTCTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACAC
AACATACGAGCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCAC
TGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTAT
TGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGG
CGGTAATACGGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAAC
CGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGA
GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCT
GCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCAATGCTCACGCTGTAGGTATCTCAGTT
CGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTAT
CGTCTTGAGTCCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTAT
GTAGGCGGTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGC
TGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGT
TTGCAAGCAGCAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGG
AACGAAAACTCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAA
GTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCG
ATCTGTCTATTTCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCC
CAGTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGA
GCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCA
GTTAATAGTTTGCGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTC
CGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTT
GTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAG
ATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGT
CAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTC
AAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCA
GCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTC
ATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAA
AATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTC

Specific sequence for clone #1257:

GCTAGCACCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGAC
AAGATGGCCCCCAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCG
CAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTC
ACTCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGG
CCCTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGA

CTGTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAA
CAGCGGTAGAGGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCCTTGAACCTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTA
CACCGGAGCAAGTCGTGGCCATTGCAAGCAATGGGGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCT
CTGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGTA
AGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTA
CACCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCT
CTGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGT
CCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGT
AAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCA
ATCGCGAACAATAATGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTT
ACACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTC
TCTGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGT
CCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGT
AAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCA
ATCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTA
CACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCT
CTGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGAATAACAATGGAGGGAAACAAGCATTGGAGACTGT
CCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGT
AAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACACCCGAACAGGTGGTCGCC
ATTGCTTCTAATGGGGGAGGACGGCCAGCCTTGGAGTCCATCGTAGCCCAATTGTCCAGGCCCGATCCCGCGTTGGCTGCGT
TAACGAATGACCATCTGGTGGCGTTGGCATGTCTTGGTGGACGACCCGCGCTCGATGCAGTCAAAAAGGGTCTGCCTCATGC
TCCCGCATTGATCAAAAGAACCAACCGGCGGATTCCCGAGAGAACTTCCCATCGAGTCGCGGGATCC

Specific sequence for clone #1258:

GCTAGCACCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGAC
AAGATGGCCCCCAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCG
CAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTC
ACTCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGG
CCCTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGA
CTGTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAA
CAGCGGTAGAGGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCCTTGAACCTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGTA
AGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTACA
CCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCT
GTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACACCCGAACAGGTGGTCGCCATTGCTTCTAATGGGGGAGGACGGCCAGCCTTGGAGTCCATCG
TAGCCCAATTGTCCAGGCCCGATCCCGCGTTGGCTGCGTTAACGAATGACCATCTGGTGGCGTTGGCATGTCTTGGTGGACG
ACCCGCGCTCGATGCAGTCAAAAAGGGTCTGCCTCATGCTCCCGCATTGATCAAAAGAACCAACCGGCGGATTCCCGAGAG
AACTTCCCATCGAGTCGCGGGATCC

Specific sequence for clone #1259:

GCTAGCACCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGAC
AAGATGGCCCCCAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCG
CAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTC
ACTCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGG
CCCTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGA
CTGTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAA
CAGCGGTAGAGGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCCTTGAACCTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAACAACAACGGCGGTAA

GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTACA
CCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCT
GTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCCAATATTGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGAACAATAATGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCCAATATTGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAATGGGGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGTCC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTCTGACACCCGAACAGGTGGTCGCCATTGCTTCTAACATCGGAGGAC
GGCCAGCCTTGGAGTCCATCGTAGCCCAATTGTCCAGGCCCGATCCCGCGTTGGCTGCGTTAACGAATGACCATCTGGTGGC
GTTGGCATGTCTTGGTGGACGACCCGCGCTCGATGCAGTCAAAAAGGGTCTGCCTCATGCTCCCGCATTGATCAAAAGAACC
AACCGGCGGATTCCCGAGAGAACTTCCCATCGAGTCGCGGGATCC

Specific sequence for clone #1260:

GCTAGCACCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGAC
AAGATGGCCCCCAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCG
CAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTC
ACTCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGG
CCCTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGA
CTGTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAA
CAGCGGTAGAGGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCCTTGAACCTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCCAATATTGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGAACAATAATGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCCAATATTGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAATGGGGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACACCCGAACAGGTGGTCGCCATTGCTTCCCACGACGGAGGACGGCCAGCCTTGGAGTCCATCG
TAGCCCAATTGTCCAGGCCCGATCCCGCGTTGGCTGCGTTAACGAATGACCATCTGGTGGCGTTGGCATGTCTTGGTGGACG
ACCCGCGCTCGATGCAGTCAAAAAGGGTCTGCCTCATGCTCCCGCATTGATCAAAAGAACCAACCGGCGGATTCCCGAGAG
AACTTCCCATCGAGTCGCGGGATCC

Specific sequence for clone #1295:

GCTAGCACCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGAC
AAGATGGCCCCCAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCG
CAACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTC
ACTCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGG
CCCTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGA
CTGTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAA
CAGCGGTAGAGGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCCTTGAACCTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGTA
AGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTACA

CCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCT
GTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTCTGACACCCGAACAGGTGGTCGCCATTGCTTCCCACGACGGAGGAC
GGCCAGCCTTGGAGTCCATCGTAGCCCAATTGTCCAGGCCCGATCCCGCGTTGGCTGCGTTAACGAATGACCATCTGGTGGC
GTTGGCATGTCTTGGTGGACGACCCGCGCTCGATGCAGTCAAAAAGGGTCTGCCTCATGCTCCCGCATTGATCAAAAGAACC
AACCGGCGGATTCCCGAGAGAACTTCCCATCGAGTCGCGGGATCC

Specific sequence for clone #1297:

GCTAGCATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAA
GATGGCCCCCAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCGCA
ACAGCAACAGGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTCAC
TCATGCGCATATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGGCC
CTGCCCGAAGCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGACT
GTGGCGGGTGAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAACA
GCGGTAGAGGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCCTTGAACCTGACCCCAGACCAGGTAGTCGCAATC
GCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTACA
CCGGAGCAAGTCGTGGCCATTGCAAGCAATGGGGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCT
GTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTCC
AACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGTAA
GCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAAT
CGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGTA
AGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTA
CACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCT
CTGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGT
CCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCTCCAATATTGGCGGTA
AGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACCCCAGACCAGGTAGTCGCAA
TCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAC
ACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTC
TGTCAAGCCCACGGGCTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTC
CAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGTA
AGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGACACCCGAACAGGTGGTCGCCA
TTGCTTCTAATGGGGGAGGACGGCCAGCCTTGGAGTCCATCGTAGCCCAATTGTCCAGGCCCGATCCCGCGTTGGCTGCGTT
AACGAATGACCATCTGGTGGCGTTGGCATGTCTTGGTGGACGACCCGCGCTCGATGCAGTCAAAAAGGGTCTGCCTCATGCT
CCCGCATTGATCAAAAGAACCAACCGGCGGATTCCCGAGAGAACTTCCCATCGAGTCGCGGGATCC

DNA sequences of plasmids expressing TALE nucleases used in this study

The DNA sequences of all six TALE nuclease expression plasmids we constructed are identical except for the region indicated by six Xs in the common sequence listed above. The specific sequences for each of the six plasmids in this variable region are given below with their associated clone names. The full expression plasmid sequence for each TALE nuclease clone can be created by substituting the specific variable sequence in place of the six Xs in the common sequence.

Supplementary Figure 5

Clone #1257:

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHG
FTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTA
VEAVHAWRNALTGAPLNLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQA
HGLTPDQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGG
GKQALETVQRLLPVLCQDHGLTPEQVVAIANNNGGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDGGKQALETVQRLLPVLC

QAHGLTPAQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIANNNGGKQALETVQRLLPVLCQDHGLTPEQVVAIASH
DGGKQALETVQRLLPVLCQAHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPDQVVAIA
NNNGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNGGGRPALESIVAQLS
RPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIEL
IEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN
KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

Clone #1258:

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHG
FTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTA
VEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIANNNGGKQALETVQRLLPVLCQA
HGLTPDQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDG
GKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQ
AHGLTPAQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPEQVVAIANNN
GGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPAQVVAIASNGGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPEQVVAIAS
NGGGRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKS
ELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQ
ADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLE
EVRRKFNNGEINF

Clone #1259:

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHG
FTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTA
VEAVHAWRNALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAH
GLTPDQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPAQVVAIANNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGK
QALETVQRLLPVLCQDHGLTPEQVVAIANNNGGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQA
HGLTPAQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIANNNGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGG
KQALETVQRLLPVLCQAHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPAQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDG
GKQALETVQRLLPVLCQAHGLTPEQVVAIASNIGGRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPA
LIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKP
DGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLT
RLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

Clone #1260:

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHG
FTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTA
VEAVHAWRNALTGAPLNLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAH
GLTPDQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPAQVVAIANNNGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGK
QALETVQRLLPVLCQDHGLTPEQVVAIANNNGGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQA
HGLTPAQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPDQVVAIANNNGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGG
KQALETVQRLLPVLCQAHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPAQVVAIASNIGGKQALETVQRLLPVLCQD
HGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDG
GRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELR
HKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADE
MQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVR
RKFNNGEINF

Clone #1295:

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHG
FTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTA
VEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIANNNGGKQALETVQRLLPVLCQA
HGLTPDQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDG
GKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQ
AHGLTPAQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNIGGKQALETVQRLLPVLCQDHGLTPEQVVAIANNN
GGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPAQVVAIASNGGGKQALETVQRLLPVL
CQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPEQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPDQVVAIAS
NGGGKQALETVQRLLPVLCQAHGLTPEQVVAIASHDGGRPALESIVAQLSRPDPALAALTNDHLVALACLGGRPALDAVKKGLP
HAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELIEIARNSTQDRILEMKVMEFFMKVYGYRGKHLGG
SRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRNKHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKA
QLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

Clone #1297:

MDYKDHDGDYKDHDIDYKDDDDKMAPKKKRKVGIHRGVPMVDLRTLGYSQQQQEKIKPKVRSTVAQHHEALVGHG
FTHAHIVALSQHPAALGTVAVKYQDMIAALPEATHEAIVGVGKQWSGARALEALLTVAGELRGPPLQLDTGQLLKIAKRGGVTA
VEAVHAWRNALTGAPLNLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNGGGKQALETVQRLLPVLCQA

HGLTPDQVVAIASNGGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPDQVVAIASHDG
GKQALETVQRLLPVLCQDHGLTPEQVVAIANNNGGKQALETVQRLLPVLCQAHGLTPDQVVAIASNGGGKQALETVQRLLPVLC
QAHGLTPAQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPDQVVAIASNGGGKQALETVQRLLPVLCQDHGLTPEQVVAIASH
DGGKQALETVQRLLPVLCQAHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQAHGLTPAQVVAIASNIGGKQALETVQRLLPV
LCQDHGLTPDQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNIGGKQALETVQRLLPVLCQAHGLTPDQVVAIAS
NGGGKQALETVQRLLPVLCQAHGLTPAQVVAIASHDGGKQALETVQRLLPVLCQDHGLTPEQVVAIASNGGGRPALESIVAQLSR
PDPALAALTNDHLVALACLGGRPALDAVKKGLPHAPALIKRTNRRIPERTSHRVAGSQLVKSELEEKKSELRHKLKYVPHEYIELI
EIARNSTQDRILEMKVMEFFMKVYGYRGKHLGGSRKPDGAIYTVGSPIDYGVIVDTKAYSGGYNLPIGQADEMQRYVEENQTRN
KHINPNEWWKVYPSSVTEFKFLFVSGHFKGNYKAQLTRLNHITNCNGAVLSVEELLIGGEMIKAGTLTLEEVRRKFNNGEINF

Full amino acid sequences of the six TALE nucleases used in this study FokI nuclease domains are underlined in each sequence.

Supplementary Figure 6

A.

```
Type I:     LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
Type II:    LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
Type III:   LTPDQVVAIASNIGGKQALETVQRLLPVLCQAHG
Type IV:    LTPAQVVAIASNIGGKQALETVQRLLPVLCQDHG
```

B.

Type I:

CTGACCCCAGACCAGGTAGTCGCAATCGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCAC
GGC

Type II:

CTTACACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTTCTCTGTCAAGCCCAC
GGG

Type III:

CTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCAC
GGT

Type IV:

TTGACGCCTGCACAAGTGGTCGCCATCGCCTCCAATATTGGCGGTAAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCAT
GGA

Comparison of amino acid and DNA sequences of Type I, II, III, and IV TALE repeats used to construct arrays

A. Amino acid sequences of Type I, II, III, and IV TALE repeats (see **Supplementary Methods**). All of the repeats shown here possess an NI RVD for binding to an A nucleotide. Amino acid residues that vary among the Type I, II, III, and IV repeats are highlighted in various colors. Sequences were obtained from Miller et al.[1]

B. DNA sequences encoding Type I, II, III, and IV TALE repeats, each bearing an NI RVD for binding to an A nucleotide, are shown.

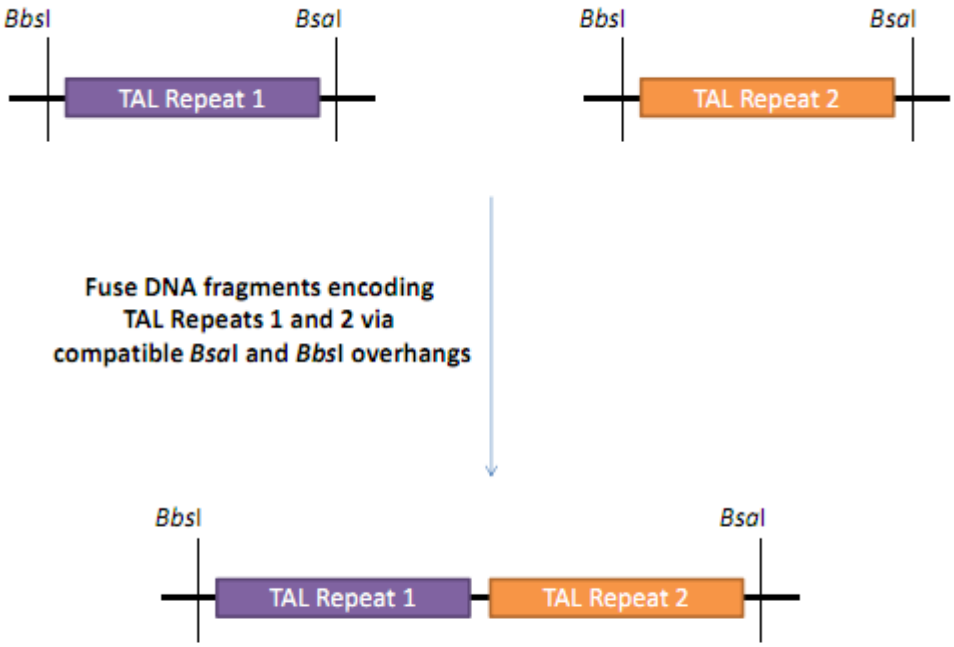
Supplementary Figure 7

TCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACGGTCAC
AGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCGGGTGT
TGGCGGGTGTCGGGGCTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCAC
CATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCCATT
CGCCATTCAGGCTGCGCAACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACG
CCAGCTGGCGAAAGGGGGATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCC
CAGTCACGACGTTGTAAAACGACGGCCAGTGAATTCGAGCTCGGTACCTCGCGAATGCATC
TAGATATCGGATCCCGGGCCCGTCGACTGCAGAGGCCTGCATGCAAGCTTGGCGTAATCAT
GGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGAGCC
GGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTT
GCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGC
CAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACT
CGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATAC
GGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAA
GGCCAGGAACCGTAAAAAGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGAC
GAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGAT
ACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTAC
CGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTA
GGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGT
TCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACAC
GACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCG
GTGCTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGAACAGTATTTGGT
ATCTGCGCTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAA
ACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAA
AAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAAC
TCACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATT
AAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATG
CTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCATAGTTGCCTGACT
CCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAATG
ATACCGCGAGAgCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAA
GGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGC
CGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCAACGTTGTTGCCATTGCTAC
AGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACGAT
CAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCG
ATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAA
TTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTGACTGGTGAGTACTCAACCAAGTC
ATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATAATA
CCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAA
CTCTCAAGGATCTTACCGCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTG
ATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGTGAGCAAAAACAGGAAGGCAAAATG
CCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTTTTCAAT
ATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAA
AAATAAACAAATAGGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAA
CCATTATTATCATGACATTAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTC

Full DNA sequence of plasmid pUC57-ΔBsaI

DNA sequences encoding the various TALE repeats shown in **Supplementary Table 1** were commercially synthesized and positioned between the unique *Xba*I and *Bam*HI sites in cloning plasmid pUC57-ΔBsaI. This plasmid encodes resistance to the antibiotic ampicillin and is identical to plasmid pUC57 except for mutation of a single base (highlighted in bold and red in the sequence above) that destroys a *Bsa*I restriction site.

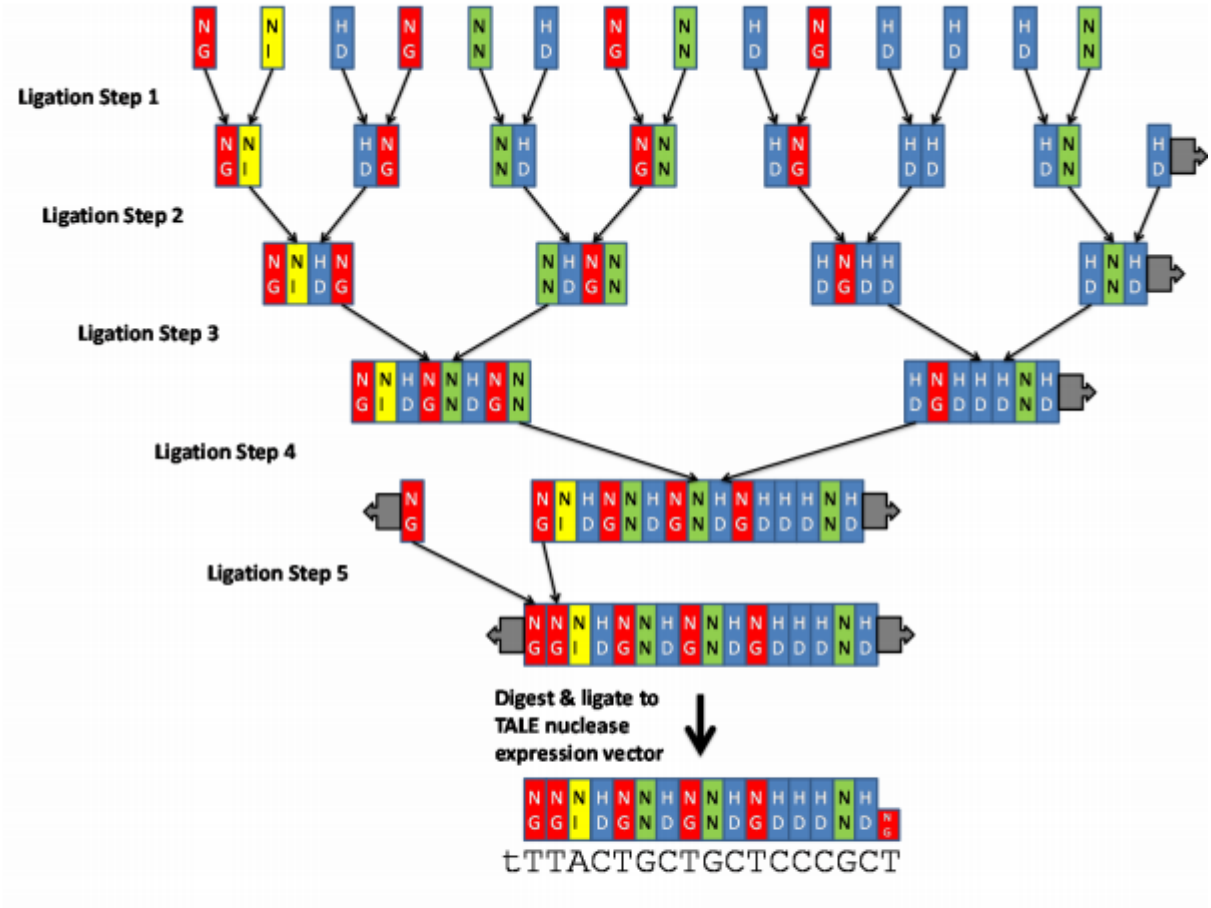
Supplementary Figure 8



Restriction enzyme-based strategy for joining together DNA sequences encoding TALE repeats

In the example shown, DNAs encoding two TALE repeats are joined together via compatible overhangs generated by digestion of the first and second DNAs with Type IIS *BsaI* and *BbsI* restriction enzymes, respectively. Following fusion of the two DNAs, flanking 5' *BbsI* and a 3' *BsaI* restriction sites remain so this DNA can be fused with others encoding additional TALE repeats. Additional details in the **Supplementary Methods**.

Supplementary Figure 9



Schematic overview of assembly strategy used to engineer TALE repeat array for TALE nuclease #1257

A series of serial ligation steps were used to assemble together the 16 TALE repeats of TALE nuclease #1257 using the restriction enzyme-based strategy detailed in **Supplementary Figure 8**. TALE repeats are shown as colored rectangles with the RVDs abbreviated as two letters. Modified sequences on the 5' and 3' ends of the DNA encoding the N- and C-terminal TALE repeats, respectively, that are required for cloning into the final TALE nuclease expression vector are illustrated as grey colored arrow boxes. Cloning into the TALE nuclease expression vector in the last ligation step creates a plasmid that expresses the TALE repeat array fused to the C-terminal 0.5 TALE repeat (shown as a smaller colored rectangle). See **Supplementary Methods** for additional details.

Supplementary Figure 10

Common Sequence:

GACGGATCGGGAGATCTCCCGATCCCCTATGGTCGACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCA
GTATCTGCTCCCTGCTTGTGTGTTGGAGGTCGCTGAGTAGTGCGCGAGCAAAATTTAAGCTACAACAAGGCAAGGCTTGACC
GACAATTGCATGAAGAATCTGCTTAGGGTTAGGCGTTTTGCGCTGCTTCGCGATGTACGGGCCAGATATACGCGTTGACATT
GATTATTGACTAGTTATTAATAGTAATCAATTACGGGGTCATTAGTTCATAGCCCATATATGGAGTTCCGCGTTACATAACTT
ACGGTAAATGGCCCGCCTGGCTGACCGCCCAACGACCCCCGCCCATTGACGTCAATAATGACGTATGTTCCCATAGTAACGC
CAATAGGGACTTTCCATTGACGTCAATGGGTGGACTATTTACGGTAAACTGCCCACTTGGCAGTACATCAAGTGTATCATAT
GCCAAGTACGCCCCCTATTGACGTCAATGACGGTAAATGGCCCGCCTGGCATTATGCCCAGTACATGACCTTATGGGACTTT
CCTACTTGGCAGTACATCTACGTATTAGTCATCGCTATTACCATGGTGATGCGGTTTTGGCAGTACATCAATGGGCGTGGATA

GCGGTTTGACTCACGGGGATTTCCAAGTCTCCACCCCATTGACGTCAATGGGAGTTTGTTTTGGCACCAAAATCAACGGGAC
TTTCCAAAATGTCGTAACAACTCCGCCCCATTGACGCAAATGGGCGGTAGGCGTGTACGGTGGGAGGTCTATATAAGCAGAG
CTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAATTAATACGACTCACTATAGGGAGACCCAAGCTGGCTA
GCaccATGGACTACAAAGACCATGACGGTGATTATAAAGATCATGACATCGATTACAAGGATGACGATGACAAGATGGCCCC
CAAGAAGAAGAGGAAGGTGGGCATTCACCGCGGGGTACCTATGGTGGACTTGAGGACACTCGGTTATTCGCAACAGCAACA
GGAGAAAATCAAGCCTAAGGTCAGGAGCACCGTCGCGCAACACCACGAGGCGCTTGTGGGGCATGGCTTCACTCATGCGCA
TATTGTCGCGCTTTCACAGCACCCTGCGGCGCTTGGGACGGTGGCTGTCAAATACCAAGATATGATTGCGGCCCTGCCCGAA
GCCACGCACGAGGCAATTGTAGGGGTCGGTAAACAGTGGTCGGGAGCGCGAGCACTTGAGGCGCTGCTGACTGTGGCGGGT
GAGCTTAGGGGGCCTCCGCTCCAGCTCGACACCGGGCAGCTGCTGAAGATCGCGAAGAGAGGGGGAGTAACAGCGGTAGA
GGCAGTGCACGCCTGGCGCAATGCGCTCACCGGGGCCCCCTTGAACAGAGACGATTAATGCGTCTCGCTGACACCCGAACA
GGTGGTCGCCATTGCTXXXXXXXXXGGAGGACGGCCAGCCTTGGAGTCCATCGTAGCCCAATTGTCCAGGCCCGATCCCGC
GTTGGCTGCGTTAACGAATGACCATCTGGTGGCGTTGGCATGTCTTGGTGGACGACCCGCGCTCGATGCAGTCAAAAAGGGT
CTGCCTCATGCTCCCGCATTGATCAAAAGAACCAACCGGCGGATTCCCGAGAGAACTTCCCATCGAGTCGCGGGATCCCAAC
TAGTCAAAAGTGAACTGGAGGAGAAGAAATCTGAACTTCGTCATAAATTGAAATATGTGCCTCATGAATATATTGAATTAAT
TGAAATTGCCAGAAATTCCACTCAGGATAGAATTCTTGAAATGAAGGTAATGGAATTTTTTATGAAAGTTTATGGATATAGA
GGTAAACATTTGGGTGGATCAAGGAAACCGGACGGAGCAATTTATACTGTCGGATCTCCTATTGATTACGGTGTGATCGTGG
ATACTAAAGCTTATAGCGGAGGTTATAATCTGCCAATTGGCCAAGCAGATGAAATGCAACGATATGTCGAAGAAAATCAAA
CACGAAACAAACATATCAACCCTAATGAATGGTGGAAAGTCTATCCATCTTCTGTAACGGAATTTAAGTTTTTATTTGTGAGT
GGTCACTTTAAAGGAAACTACAAAGCTCAGCTTACACGATTAAATCATATCACTAATTGTAATGGAGCTGTTCTTAGTGTAG
AAGAGCTTTTAATTGGTGGAGAAATGATTAAAGCCGGCACATTAACCTTAGAGGAAGTCAGACGGAAATTTAATAACGGCG
AGATAAACTTTTAAGGGCCCTTCGAAGGTAAGCCTATCCCTAACCCTCTCCTCGGTCTCGATTCTACGCGTACCGGTCATCAT
CACCATCACCATTGAGTTTAAACCCGCTGATCAGCCTCGACTGTGCCTTCTAGTTGCCAGCCATCTGTTGTTTGCCCCTCCCCC
GTGCCTTCCTTGACCCTGGAAGGTGCCACTCCCACTGTCCTTTCCTAATAAAATGAGGAAATTGCATCGCATTGTCTGAGTAG
GTGTCATTCTATTCTGGGGGGTGGGGTGGGGCAGGACAGCAAGGGGGAGGATTGGGAAGACAATAGCAGGCATGCTGGGG
ATGCGGTGGGCTCTATGGCTTCTGAGGCGGAAAGAACCAGCTGGGGCTCTAGGGGGTATCCCCACGCGCCCTGTAGCGGCG
CATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTC
TTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGCATCCCTTTAGGGTTCCGATTTAGTGCT
TTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCC
CTTTGACGTTGGAGTCCACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTT
TTGATTTATAAGGGATTTTGGGGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTAATTC
TGTGGAATGTGTGTCAGTTAGGGTGTGGAAAGTCCCCAGGCTCCCCAGGCTCCCCAGGCAGGCAGAAGTATGCAAAGCATGCATCTCAATT
AGTCAGCAACCAGGTGTGGAAAGTCCCCAGGCTCCCCAGCAGGCAGAAGTATGCAAAGCATGCATCTCAATTAGTCAGCAA
CCATAGTCCCGCCCCTAACTCCGCCCATCCCGCCCCTAACTCCGCCCAGTTCCGCCCATTCTCCGCCCCATGGCTGACTAATT
TTTTTTATTTATGCAGAGGCCGAGGCCGCCTCTGCCTCTGAGCTATTCCAGAAGTAGTGAGGAGGCTTTTTTGGAGGCCTAGG
CTTTTGCAAAAAGCTCCCGGGAGCTTGTATATCCATTTTCGGATCTGATCAGCACGTGTTGACAATTAATCATCGGCATAGTA
TATCGGCATAGTATAATACGACAAGGTGAGGAACTAAACCATGGCCAAGCCTTTGTCTCAAGAAGAATCCACCCTCATTGAA
AGAGCAACGGCTACAATCAACAGCATCCCCATCTCTGAAGACTACAGCGTCGCCAGCGCAGCTCTCTCTAGCGACGGCCGCA
TCTTCACTGGTGTCAATGTATATCATTTTACTGGGGGACCTTGTGCAGAACTCGTGGTGCTGGGCACTGCTGCTGCTGCGGCA
GCTGGCAACCTGACTTGTATCGTCGCGATCGGAAATGAGAACAGGGGCATCTTGAGCCCCTGCGGACGGTGTCGACAGGTG
CTTCTCGATCTGCATCCTGGGATCAAAGCGATAGTGAAGGACAGTGATGGACAGCCGACGGCAGTTGGGATTCGTGAATTGC
TGCCCTCTGGTTATGTGTGGGAGGGCTAAGCACTTCGTGGCCGAGGAGCAGGACTGACACGTGCTACGAGATTTCGATTCCA
CCGCCGCCTTCTATGAAAGGTTGGGCTTCGGAATCGTTTTCCGGGACGCCGGCTGGATGATCCTCCAGCGCGGGGATCTCAT
GCTGGAGTTCTTCGCCCACCCCAACTTGTTTATTGCAGCTTATAATGGTTACAAATAAAGCAATAGCATCACAAATTTCACAA
ATAAAGCATTTTTTTCACTGCATTCTAGTTGTGGTTTGTCCAAACTCATCAATGTATCTTATCATGTCTGTATACCGTCGACCT
CTAGCTAGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTGAAATTGTTATCCGCTCACAATTCCACACAACATACGA
GCCGGAAGCATAAAGTGTAAAGCCTGGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTT
TCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGAGGCGGTTTGCGTATTGGGCGCTC
TTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGTTCGGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATAC
GGTTATCCACAGAATCAGGGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCGTAAAAAG
GCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGCTCAAGTCAGAGGTGGCGAA
ACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACC
GGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCAATGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGT
CGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGT
CCAACCCGGTAAGACACGACTTATCGCCACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTG
CTACAGAGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATCTGCGCTCTGCTGAAGCCAGT
TACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAACAAACCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAG
CAGATTACGCGCAGAAAAAAAGGATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAAACT
CACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTTTAAATTAAAAATGAAGTTTTAAATCA
ATCTAAAGTATATATGAGTAAACTTGGTCTGACAGTTACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATT
TCGTTCATCCATAGTTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCCCCAGTGCTGCAA
TGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAAACCAGCCAGCCGGAAGGGCCGAGCGCAGAAGTG
GTCCTGCAACTTTATCCGCCTCCATCCAGTCTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTG
CGCAACGTTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATTCAGCTCCGGTTCCCAACG
ATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAGCGGTTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAG
TTGGCCGCAGTGTTATCACTCATGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTCTGTG
ACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTTGCTCTTGCCCGGCGTCAATACGGGATA
ATACCGCGCCACATAGCAGAACTTTAAAAGTGCTCATCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACC
GCTGTTGAGATCCAGTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAGCGTTTCTGGGT

GAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGCGACACGGAAATGTTGAATACTCATACTCTTCCTTT
TTCAATATTATTGAAGCATTTATCAGGGTTATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATA
GGGGTTCCGCGCACATTTCCCCGAAAAGTGCCACCTGACGTC

DNA sequences of plasmids for expressing TALE nucleases

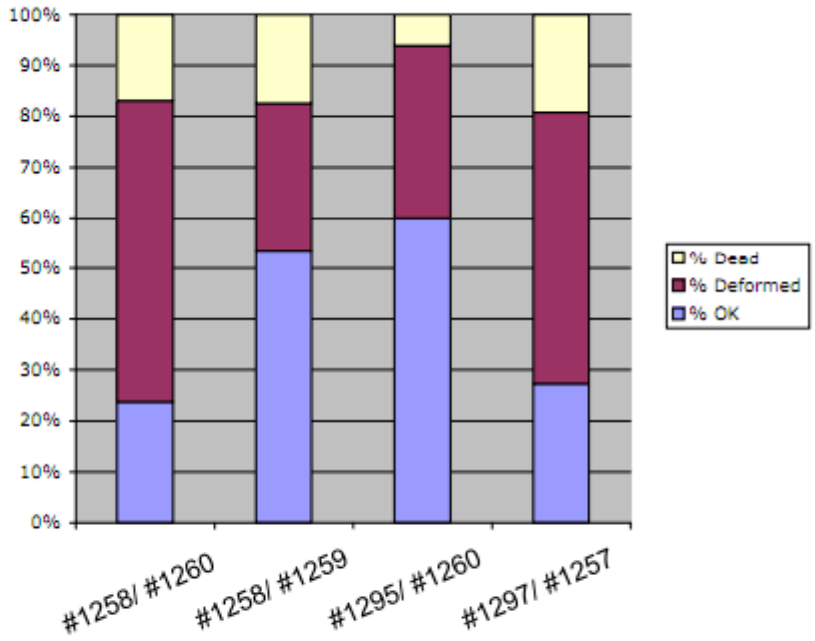
As described in Supplementary Methods, we constructed three different plasmids into which DNA fragments encoding assembled TALE arrays can be cloned to generate TALE nuclease expression vectors. The three plasmids differ in the identity of the C-terminal 0.5 TALE repeat.

All plasmids share the common sequence above and differ at just nine nucleotide positions marked as XXXXXXXXX (underlined and bold). The sequence of these 9 bps and plasmid names are also shown below for vectors that have the NI, HD, and NG RVDs as the C-terminal

0.5 TALE repeat.

Plasmid name	Sequence of variable 9 bps	RVD of C-terminal 0.5 TALE repeat
pJDS70	TCTAACATC	NI (for binding to an A nucleotide)
pJDS71	TCCCACGAC	HD (for binding to a C nucleotide)
pJDS78	TCTAATGGG	NG (for binding to a T nucleotide)

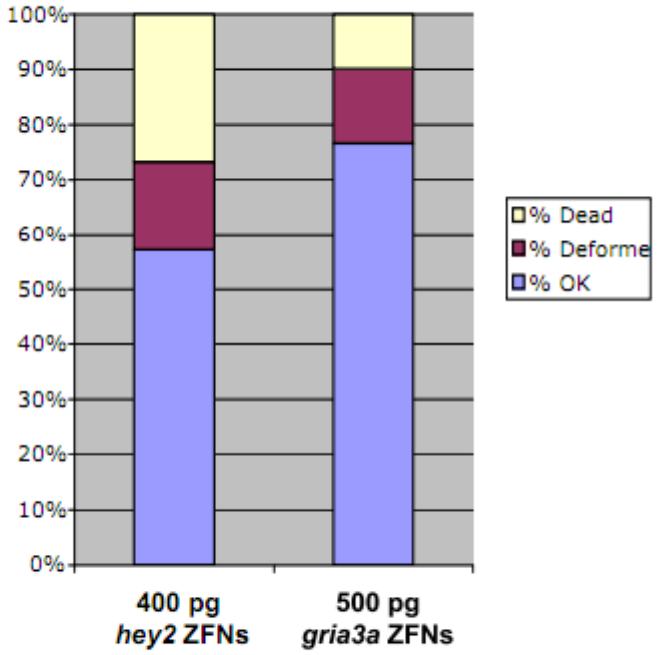
Supplementary Figure 11



Toxicity of TALE nucleases in zebrafish embryos

Percentages of dead, deformed and normal ("OK") embryos one day post-injection are shown.

Supplementary Figure 12



Toxicity of zinc finger nucleases in zebrafish embryos.

Percentages of dead, deformed and normal ("OK") embryos one day post-injection are shown.

Supplementary Table 1

Plasmid ID #	Type	DNA Target / RVD	DNA Sequence	Amino Acid Sequence
A	Type I	A/NI	TCTAGAGAAGACAACTGACCCCAGACCAGGTAGTCGCAATCGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAAGAGACCGGATCC	LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
B	Type I	C/HD	TCTAGAGAAGACAACTGACCCCAGACCAGGTAGTCGCAATCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAAGAGACCGGATCC	LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
C	Type I	G/NN	TCTAGAGAAGACAACTGACCCCAGACCAGGTAGTCGCAATCGCGAACAATAATGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAAGAGACCGGATCC	LTPDQVVAIANNGGKQALETVQRLLPVLCQDHG
D	Type I	T/NG	TCTAGAGAAGACAACTGACCCCAGACCAGGTAGTCGCAATCGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAAGAGACCGGATCC	LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
E	Type II	A/NI	TCTAGAGAAGACAACTTACACCGGAGCAAGTCGTGGCCATTGCAAGCAACATCGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCCAGTCTCTGTCAAGCCCACGGGCTGAAGAGACCGGATCC	LTPEQVVAIASNIGGKQALETVQRLLPVLCQAHG
F	Type II	C/HD	TCTAGAGAAGACAACTTACACCGGAGCAAGTCGTGGCCATTGCATCCCACGACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCCAGTCTCTGTCAAGCCCACGGGCTGAAGAGACCGGATCC	LTPEQVVAIASHDGGKQALETVQRLLPVLCQAHG
G	Type II	G/NN	TCTAGAGAAGACAACTTACACCGGAGCAAGTCGTGGCCATTGCAAATAATAACGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCCAGTCTCTGTCAAGCCCACGGGCTGAAGAGACCGGATCC	LTPEQVVAIANNGGKQALETVQRLLPVLCQAHG
H	Type II	T/NG	TCTAGAGAAGACAACTTACACCGGAGCAAGTCGTGGCCATTGCAAGCAATGGGGGTGGCAAACAGGCTCTTGAGACGGTTCAGAGACTTCTCCCAGTCTCTGTCAAGCCCACGGGCTGAAGAGACCGGATCC	LTPEQVVAIASNGGGKQALETVQRLLPVLCQAHG
I	Type III	A/NI	TCTAGAGAAGACAACTGACTCCCGATCAAGTTGTAGCGATTGCGTCGAACATTGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTGAAGAGACCGGATCC	LTPDQVVAIASNIGGKQALETVQRLLPVLCQAHG
J	Type III	C/HD	TCTAGAGAAGACAACTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTGAAGAGACCGGATCC	LTPDQVVAIASHDGGKQALETVQRLLPVLCQAHG
K	Type III	G/NN	TCTAGAGAAGACAACTGACTCCCGATCAAGTTGTAGCGATTGCGAATAACAATGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGAAGAGACCGGATCC	LTPDQVVAIANNGGKQALETVQRLLPVLCQAHG
L	Type III	T/NG	TCTAGAGAAGACAACTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTTTGAAGAGACCGGATCC	LTPDQVVAIASNGGGKQALETVQRLLPVLCQAHG
M	Type IV	A/NI	TCTAGAGAAGACAATTGACGCCTGCACAAGTGGTCGCCATCGCCTCCAATATTGGCGGTAAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGAAGAGACCGGATCC	LTPAQVVAIASNIGGKQALETVQRLLPVLCQDHG
N	Type IV	C/HD	TCTAGAGAAGACAATTGACGCCTGCACAAGTGGTCGCCATCGCCAGCCATGATGGCGGTAAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGAAGAGACCGGATCC	LTPAQVVAIASHDGGKQALETVQRLLPVLCQDHG
O	Type IV	G/NN	TCTAGAGAAGACAATTGACGCCTGCACAAGTGGTCGCCATCGCCAACAACAACGGCGGTAAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGAAGAGACCGGATCC	LTPAQVVAIANNGGKQALETVQRLLPVLCQDHG
P	Type IV	T/NG	TCTAGAGAAGACAATTGACGCCTGCACAAGTGGTCGCCATCGCCTCGAATGGCGGCGGTAAGCAGGCGCTGGAAACAGTACAGCGCCTGCTGCCTGTACTGTGCCAGGATCATGGACTGAAGAGACCGGATCC	LTPAQVVAIASNGGGKQALETVQRLLPVLCQDHG
Q	Type I (N-terminal)	A/NI	TCTAGAGAAGACAAGAACCTGACCCCAGACCAGGTAGTCGCAATCGCGTCGAACATTGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAAGAGACCGGATCC	LTPDQVVAIASNIGGKQALETVQRLLPVLCQDHG
R	Type I (N-terminal)	C/HD	TCTAGAGAAGACAAGAACCTGACCCCAGACCAGGTAGTCGCAATCGCGTCACATGACGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAAGAGACCGGATCC	LTPDQVVAIASHDGGKQALETVQRLLPVLCQDHG
S	Type I (N-terminal)	T/NG	TCTAGAGAAGACAAGAACCTGACCCCAGACCAGGTAGTCGCAATCGCGTCAAACGGAGGGGGAAAGCAAGCCCTGGAAACCGTGCAAAGGTTGTTGCCGGTCCTTTGTCAAGACCACGGCCTTAAGAGACCGGATCC	LTPDQVVAIASNGGGKQALETVQRLLPVLCQDHG
T	Type III (C-terminal)	C/HD	TCTAGAGAAGACAACTGACTCCCGATCAAGTTGTAGCGATTGCGTCGCATGACGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTCTGAAGAGACCGGATCC	LTPDQVVAIASHDGGKQALETVQRLLPVLCQAHG
U	Type III (C-terminal)	T/NG	TCTAGAGAAGACAACTGACTCCCGATCAAGTTGTAGCGATTGCGTCCAACGGTGGAGGGAAACAAGCATTGGAGACTGTCCAACGGCTCCTTCCCGTGTTGTGTCAAGCCCACGGTCTGAAGAGACCGGATCC	LTPDQVVAIASNGGGKQALETVQRLLPVLCQAHG

DNA and amino acid sequences of individual TALE repeats used to construct TALE nucleases

Each of the plasmids encoding the various TALE repeats in the table was commercially synthesized to have the *Xba*I-*BamH*I DNA sequence shown in the table positioned between the unique *Xba*I and *BamH*I sites in plasmid pUC57-ΔBsaI. The amino acid sequence of the individual TALE repeats are also shown. The full sequence of pUC57-ΔBsaI is given in **Supplementary Figure 7**.

Supplementary Discussion

Criteria used by ZiFiT Targeter to pick potential TALEN cleavage sites

Users input DNA sequences of interest into ZiFiT Targeter with a single target nucleotide bracketed. The software then attempts to identify potential TALEN target sites where the bracketed nucleotide falls within the "spacer" sequence between the half-sites bound by TALEN monomers. As the range of sequences that can be targeted using TALENs is very large, ZiFiT Targeter uses the iterative approach described below to identify as many as five potential target sites.

ZiFiT Targeter first truncates the query sequence to 101 base pairs (50 base pairs flanking the bracketed nucleotide). This sequence is then scanned for all potential TALEN target sites in which the bracketed nucleotide falls within the spacer sequence between the half sites and in which half-sites are 16 to 18 base pairs in length (including the conserved 5' T and the 3'-most nucleotide bound by the 0.5 TALE repeat array) and spacer sequences are 16 to 18 base pairs in length. Full target sites chosen are also filtered to ensure that they differ from each other by at least six base pairs (three base pairs in each half-site). This list is then rank-ordered based on the total length of the target site (i.e.—the sum of the lengths of the target half-sites) and the top five sites are reported as "First Tier" sites.

If ZiFiT Targeter fails to find five "First Tier" sites, it will identify "Second Tier" sites by performing the same search but loosening constraints to allow spacer sequences 13 to 23 base pairs in length. As with First Tier sites, these sites are filtered to ensure that they differ from each other by at least six base pairs and then rank-ordered based on the total length of the full site. If there are multiple target sites of the same length, these subsets of sites are secondarily ranked according to the degree of similarity between the spacer length and an ideal spacer length of 17 base pairs. For example, a spacer length of 15 base pairs (difference of two relative to 17) is more highly ranked than a spacer length of 20 base pairs (difference of three relative to 17). A total of five First Tier and Second Tier target sites are then displayed.

The criteria used above for selecting target sites are based on published data from Miller et al. and some of our own unpublished experience.

If ZiFiT Targeter fails to identify a total of five First Tier and Second Tier sites, the program provides users with the option to relax search criteria further to allow identification of potential TALEN target sites with half-sites 12 to 22 base pairs in length (including the conserved 5’ T) and spacer sequences 13 to 23 base pairs in length. This can be done by clicking the box for “Relax Search Criteria” underneath the sequence input box and clicking the Search button again. This relaxed search will still filter sites to ensure that they differ from each other by at least six base pairs. Users can expand the list of potential sites even further by eliminating the filter for similar sites. This can be accomplished by unclicking the box for “Mask Redundant Sites” and then clicking the Search button again.

APPENDIX D – CURRICULUM VITAE

CAREER OBJECTIVE

- Research scientist in the field of Bioinformatics or Computational Biology.

EDUCATION

- B.S. in Computer Science
  2005
  Oklahoma City University
- Ph.D. in Bioinformatics and Computational Biology
  2007 – Present
  Iowa State University

EXPERIENCE

- Java Programmer
  2005 – 2007
  Emergency Physicians Billing Services

POSITION HELD

- Graduate Research Assistant
  2007 – Present
  Bioinformatics and Computation Biology Program

- Teaching Assistant
  Fall, 2008 *Biol 313L Principles of Genetics*
  Spring, 2009 *Biol 313L Principles of Genetics*
  Fall 2009 *BCB 444/544 Introduction to Bioinformatics*
  Genetics Development and Cell Biology Department

AWARDS

- Professional Advancement Grant, ISU – Fall 2008
- Professional Advancement Grant, ISU – Fall 2009
- Fung Travel Award, GDCB, ISU – Fall 2009
- Fung Travel Award, GDCB, ISU – Spring 2010

PUBLICATIONS

An asterisk (*) denotes publications included in this dissertation

1. * Foley, J.E., Yeh, J-R.J., Maeder, M.L., **Reyon, D.**, Sander, J.D., Peterson, R.T., Joung, J.K. Rapid Mutation of Endogenous Zebrafish Genes Using Zinc Finger Nucleases Made by Oligomerized Pool ENgineering (OPEN). PLoS One 4(2): e4348. 2009
2. * Sander, J.D., **Reyon, D**., Maeder, M.L., Foley, J.E., Beganny, S.T., Fu, F., Voytas, D.F., Joung, J.K., Dobbs, D. Predicting Oligomerized Pool ENgineering (OPEN) success for zinc finger target sites sequences. BMC Bioinformatics. 2010 Nov 2. **(Co-first author)**
3. Su, C.C., Yang, F., Long, F., **Reyon, D.**, Routh, M.D., Kuo, D.W., Mokhtari, A.K., Van Ornam, J.D., Rabe, K.L., Hoy, J.A., Lee, Y.J., Rajashankar, K.R., Yu, E.W. Crystal Structure of the Membrane Fusion Protein CusB from Escherichia coli. Journal of Molecular Biology, 2009 Oct 23.
4. Zhang, F., Maeder, M., Unger-Wallace, E., Hoshaw, J., Reyon, D., Christian, M., Li, X., Pierick, C., Dobbs, D., Peterson, T., Joung. J. K., Voytas, D. High frequency targeted mutagenesis in Arabidopsis thaliana using zinc finger nucleases. Proceedings of the National Academy of Sciences, 2010 March 24.
5. * Sander, J.D., Maeder, M.L., **Reyon, D.**, Voytas, D.F., Joung, J.K., Dobbs, D., ZiFiT (Zinc Finger Targeter): an updated zinc finger engineering tool. Nucleic Acids Research, 2010 April 30.
6. * Sander, JD., Dahlborg, E., Goodwin, M., Cade, L., Zheng, F., Cifuentes, D., Curtin, SJ., Thibodeau-Beganny, S., Qi, Y., Pierick, CJ., Hoffman, E., Maeder, ML., Khayter, C., **Reyon, D.**, Dobbs, D., Stupar, RM., Giraldez, A., Voytas, DF., Peterson, RT., Yeh, JRJ., Joung., JK. Context-Dependent Assembly (CoDA): A Highly Efficient, Selection-Free Method for Engineering Zinc Finger Nucleases. Nature Methods, 2010 December 12.
7. *** Reyon,D.**, Kirkpatrick,JR., Sander,JD., Zhang,F., Voytas,DF., Joung.JK., Dobbs,D., Coffman,CR. ZFNGenome: A comprehensive resource for locating zinc finger nuclease target sites in model organisms. BMC Genomics. 2011 Jan. 28.
8. Curtin, S.J., Zhang, F., Sander, J.D., Haun, W.J., Starker, C., Baltes, N.J., **Reyon, D.,** Dahlborg, E.J., Goodwin, M.J., Coffman, A.P., Dobbs, D, Joung, J.K., Voytas, D.F., Stupar, R.M., Targeted mutagenesis of duplicated genes in soybean with zinc-finger nucleases. Plant Physiology. 2011 Apr 4.
9. * Sander, J.D., Cade, L., Khayter, C., **Reyon, D.,** Peterson, R.T., Joung, J.K., Yeh, J.R., Targeted gene disruption in somatic zebrafish cells using engineered TALENs. Nature Biotechnology. 2011 Aug 5.
10. * Sander, J.D., **Reyon, D.,** Khayter, C., Joung, J.K., User-friendly protocol and software for rapid engineering of designer TALE Nucleases (TALENs). Nature Protocols (Under review). **(Co-first author).**

ORAL AND POSTER PRESENTATIONS AT CONFERENCES

Presenting author is underlined

11. **Reyon, D.,** Lewis, B., Lee, J.H., Gleeson, C., Hamilton, M., Caragea, C., Towfic, F., Terribilini, M., Honavar, V., Dobbs, D. Combining Structural Modeling, Evolutionary Information, and Machine Learning to Improve Prediction of Nucleic Acid Binding Sites in Telomerase. 5th Annual Rocky Mountain Bioinformatics Conference (Rocky '07). Snowmass, CO. **Oral Presentation and Poster.**

12. **Reyon, D.,** Sucaet, Y., Towfic, F., Boyken, S., Farnham, R., Nikolova, O., Hemert, J.V. Eighteen months and counting: A status update of activities in Iowa State University's BCBLab initiative. New Mexico Bioinformatics Symposium (NMBIS' 08). Santa Fe, NM. **Oral Presentation *(one of only 2 student abstracts chosen)* and Poster**

13. **Reyon, D.,** Sander, J.D., Joung, J.K., Voytas, D.F., Dobbs, D. ZFNGenome: a database tool for analysis of OPEN-generated Zinc Finger Nuclease target sites in mice and men (and more) Conference on Genomic Engineering 2009, Minneapolis, MN. **Poster Presentation**

14. Boyken, S., **Reyon, D.,** Lewis, B., Caragea, C., Lee, J.H., Terribilini, M., Honavar, V., Dobbs, D., Influence of Secondary Structure on Prediction of RNA Binding Sites in Protein. New Mexico Bioinformatics Symposium (NMBIS' 08). Santa Fe, NM. **Poster.**

15. Terribilini, M., Caragea, C., **Reyon, D.,** Lewis, B., Xue, L., Sander, J., Lee, J.H., Jernigan, R., Honavar, V., Rajan, K., Dobbs, D., Comparing Sequence and Structure based Classifiers for Predicting RNA Binding Sites in Specific Families of RNA Binding Proteins. Intelligent Systems for Molecular Biology (ISMB '08). Toronto, Canada. **Poster Presentation**

16. **Lewis, B.,** Kurcinski, M., Reyon, D., Lee, J.H., Honavar, V., Jernigan, R., Kolinski, A.,Kloczkowski, A., Dobbs, D., Combining Predictions of Protein Structure and Protein RNA Interactions to Model Human Telomerase Structure. Intelligent Systems for Molecular Biology (ISMB '08). Toronto, Canada. **Poster.**

17. **Reyon, D.,** Kirkpatrick, J.R., Sander, J.D., Joung, J.K., Voytas, D.F., Coffman ,C.R., & Dobbs, D., ZFNGenome: A GBrowse-based tool for identifying Zinc Finger Nuclease target sites in model organisms. Genome Engineering: Research & Therapeutic Applications (FASEB '10). Steamboat Springs, Colorado. **Poster.**

18. Sander, J.D., **Reyon, D.,** Kuzmickas, R., Khayter, C., Morgan, M.L., Joung, J.K., Further Improvements and Applications of Context-Dependent Assembly (CoDA): A Publicly Available, Selection-Free Method for Engineering Zinc Finger Nucleases. American Society of Gene & Cell Therapy (ASGCT '11). Seattle, Washington. **Poster.**